М.В. Абакумов, Н.В. Соснин

ЛЕКЦИИ ПО ЧИСЛЕННЫМ МЕТОДАМ

МОСКВА – 2022

УДК 519.6
ББК 22.192
А13

Абакумов М.В., Соснин Н.В.

А13 **Лекции по численным методам.** — М.: МАКС Пресс, 2022. — 158с.

ISBN 978-5-317-06770-0

Методическое пособие отражает содержание лекционного курса «Численные методы», читаемого студентам факультета вычислительной математики и кибернетики МГУ имени М.В.Ломоносова.

Излагаются прямые и итерационные методы решения систем линейных алгебраических уравнений. Обсуждаются некоторые методы решения задач на собственные значения. Рассматриваются численные методы решения нелинейных уравнений. Затрагиваются отдельные приемы интерполирования и приближения функций. Приводятся методы численного решения задачи Коши и краевой задачи для обыкновенного дифференциального уравнения. Формулируются основные методы построения, исследования и численной реализации разностных схем краевых задач для дифференциальных уравнений в частных производных. Содержится набор упражнений, способствующий усвоению излагаемого материала.

Пособие рассчитано на студентов старших курсов, магистрантов и аспирантов, специализирующихся в области прикладной математики.

УДК 519.6
ББК 22.192

ISBN 978-5-317-06770-0

Оглавление

Предисловие **7**

1 Методы решения систем линейных алгебраических уравнений **8**

1.1 Прямые методы решения систем линейных алгебраических уравнений 8

1.1.1 Метод квадратного корня (метод Холецкого) 11

1.1.2 Модифицированный метод квадратного корня 12

1.2 Итерационные методы решения систем линейных алгебраических уравнений 14

1.2.1 Линейные одношаговые итерационные методы 15

1.2.2 Примеры одношаговых линейных итерационных методов 17

1.3 Условия сходимости одношаговых стационарных итерационных методов 19

1.3.1 Необходимые и достаточные условия сходимости 19

1.3.2 Оценка скорости сходимости одношаговых стационарных методов 24

1.3.3 Модельная задача 27

1.4 Попеременно–треугольный итерационный метод 30

1.4.1 Алгебраическая теория 30

1.4.2 Чебышевский набор итерационных параметров 34

1.4.3 Попеременно–треугольный итерационный метод с упорядоченным набором чебышевских параметров 38

1.5 Итерационные методы вариационного типа 39

1.5.1 Одношаговые итерационные методы вариационного типа 39

1.5.2 Примеры одношаговых итерационных методов вариационного типа 44

1.6 Методы сопряженных направлений 46
1.6.1 Метод сопряженных градиентов 46
1.6.2 Метод Крейга 55
1.6.3 Симметризованные сопряжённые градиенты 55

2 Задачи на собственные значения 57
2.1 Поиск собственных значений методом вращений 57
2.2 Степенной метод поиска собственных значений 61
2.2.1 Поиск максимального и минимального собственного значения 63
2.2.2 Поиск собственного значения ближайшего к заданному числу 64
2.3 Метод обратных итераций 64

3 Численные методы решения нелинейных уравнений 66
3.1 Методы разделения корней 66
3.2 Примеры итерационных методов вычисления корней 67
3.2.1 Метод простой итерации 67
3.2.2 Метод Ньютона 69
3.2.3 Модифицированный метод Ньютона 70
3.2.4 Метод секущих 70
3.3 Сходимость метода простой итерации 72
3.4 Метод Эйткена 74
3.5 Сходимость метода Ньютона 75
3.6 Решение систем нелинейных уравнений 77
3.7 Примеры 78
3.7.1 Решение нелинейного уравнения 78
3.7.2 Решение системы нелинейных уравнений 81

4 Интерполирование и приближение функций 83
4.1 Интерполирование алгебраическими многочленами 83
4.2 Интерполирование сплайнами 87
4.2.1 Интерполирование кубическими сплайнами 87
4.2.2 Сходимость процесса интерполирования кубическими сплайнами 89
4.3 Наилучшее приближение в гильбертовом пространстве 94

5 Численное решение задачи Коши для обыкновенного дифференциального уравнения 100

5.1 Методы Рунге-Кутта 102

5.2 Многошаговые методы 108

5.3 Методы Адамса и Гира 111

5.3.1 Многошаговые методы Адамса 111

5.3.2 Многошаговые методы Гира 114

5.4 Устойчивость численных методов решения задачи Коши 116

5.5 Численное решение задачи Коши для системы обыкновенных дифференциальных уравнений 119

6 Численное решение краевой задачи для обыкновенного дифференциального уравнения 121

6.1 Разностная схема 121

6.1.1 Интегро-интерполяционный метод построения разностной схемы 122

6.1.2 Метод аппроксимации квадратичного функционала . . . 124

6.2 Элементы теории разностных схем 126

6.2.1 Решение разностной схемы 126

6.2.2 Порядок аппроксимации 127

6.2.3 Устойчивость разностной схемы 128

6.2.4 Сходимость решения разностной схемы 129

7 Численное решение краевых задач для дифференциальных уравнений в частных производных 132

7.1 Разностные схемы для уравнений параболического типа 132

7.1.1 Явная разностная схема для уравнения теплопроводности 133

7.1.2 Неявная разностная схема для уравнения теплопроводности 140

7.1.3 Разностная схема с весами для уравнения теплопроводности 143

7.1.4 Разностная схема для уравнения теплопроводности с переменными коэффициентами 148

7.1.5 Разностная схема для нелинейного уравнения теплопроводности 150

7.2 Разностная схема для уравнения колебаний 151

7.3 Разностная аппроксимация задачи Дирихле для уравнения Пуассона 153

Предисловие

В основу данного учебного пособия положены материалы лекций, читавшихся авторами студентам третьего курса факультета ВМК МГУ имени М.В.Ломоносова.

При написании данного учебного пособия авторы преследовали цель упрощения изложения, сокращения объема книги. В книге содержатся разделы, входящие в план обязательного учебного курса «Численные методы».

Изложение начинается с рассмотрения методов решения систем линейных алгебраических уравнений. Кратко обсуждаются методы вычисления собственных значений матриц. Приводятся алгоритмы методов численного решения нелинейных уравнений. Рассмотрены некоторые вопросы интерполирования и приближения функций. Формулируются эффективные методы численного решения начальной задачи для обыкновенного дифференциального уравнения. Представлены разностные методы решения краевых задач математической физики. Обсуждаются приемы построения разностных схем для различных типов уравнений. Уделяется внимание исследованию устойчивости, сходимости и методам решения сеточных уравнений.

Считаем, что данное учебное пособие окажется полезным для студентов и преподавателей, интересующихся численными методами решения математических задач.

М.В. Абакумов, Н.В. Соснин

Глава 1

Методы решения систем линейных алгебраических уравнений

В данной главе будем рассматривать знакомую, например, из линейной алгебры, задачу — требуется решить систему линейных алгебраических уравнений, записанных в виде матричного уравнения $Ay = f$, где $A = (a_{ij})$ — заданная квадратная матрица размерности $n \times n$ $(i, j = 1, 2, \ldots, n)$, $y = (y_1, y_2, \ldots, y_n)^T$ — вектор-столбец неизвестных, $f = (f_1, f_2, \ldots, f_n)^T$ — заданный вектор–столбец правых частей. Все параметры a_{ij} и f_i — вещественные числа.

Здесь и далее будем предполагать, что у рассматриваемых задач решение существует и единственно. В данном случае, условие, что определитель матрицы A не равен нулю $(\det A \neq 0)$, гарантирует существование и единственность решения системы линейных алгебраических уравнений [5].

1.1 Прямые методы решения систем линейных алгебраических уравнений

Рассмотрим несколько прямых методов для решения системы

$$Ay = f. \tag{1.1}$$

Известным, широко используемым методом решения систем линейных алгебраических уравнений с невырожденной матрицей $(\det A \neq 0)$, является *метод Гаусса*[1]. Применять метод Гаусса можно тогда и только тогда, когда

[1] Впервые описан К. Гауссом в 1849г.

все угловые миноры матрицы A отличны от нуля ($\det A$ является угловым минором $n-$ого порядка). Кроме того, метод Гаусса в классическом виде является неустойчивым методом [10]. Для того, чтобы обойти эти ограничения, на практике используют расчетные формулы метода Гаусса в сочетании с некоторой схемой выбора главного элемента [3].

Метод Гаусса является прямым методом решения систем линейных алгебраических уравнений. К этой группе относят методы, в которых получают искомый вектор y за конечное число арифметических операций.

Основным показателем при оценки эффективности конкретного метода является количество арифметических операций, необходимых для вычисления y. Общее число операций умножения и деления, более длительных по времени их реализации на вычислительной технике по сравнению с операциями сложения и вычитания, в методе Гаусса равно $n^3/3 + O(n^2)$. Объем числовой информации, которую необходимо хранить при реализации метода Гаусса, составляет $O(n^2)$. Для современных персональных компьютеров при $n \approx 10^4$ число арифметических операций и объем памяти, требующийся для реализации метода Гаусса, вполне приемлемы. Поэтому использование стандартных программ, реализующих метод Гаусса с выбором главного элемента, эффективно при решении систем линейных алгебраических уравнений с числом уравнений $n \leqslant 10^4$. Одной из таких, проверенных вычислительной практикой стандартных программ, имеющейся в свободном доступе, является программа Y12M. Теоретическое обоснование, использованного в программе Y12M расчетного алгоритма приведено в [9].

Возможны ситуации, когда возникает потребность в использовании для решения систем линейных алгебраических уравнений методов более экономичных по затратам, чем метод Гаусса. Конкурируют по трудоемкости с методом Гаусса только методы, в которых явно учитывается специфика матрицы A, то есть методы пригодные для некоторых частных видов систем линейных алгебраических уравнений. Примером может служить система линейных алгебраических уравнений с трехдиагональной матрицей A. Оптимальным методом решения такой системы является, как известно, метод прогонки [3], для реализации которого требуется $O(n)$ арифметических операций умножения и деления.

Симметричность элементов матрицы A относительно главной диагонали ($A^T = A$) является частным свойством матрицы. Положительность матрицы также является частным свойством матрицы. Напомним, что матрица A называется *положительной* ($A > 0$), если для любого вектора $y \neq 0$ скалярное произведение $(Ay, y) > 0$. Положительность эквивалентна положительности всех угловых миноров матрицы A (критерий Сильвестра) или, для ($A^T = A$), положительности всех собственных чисел $\lambda(A)$ матрицы A (см. [1]). Заметим,

что любой из указанных критериев положительности матрицы A затруднительно проверить для матриц большой размерности.

Отметим одно простое и удобное для проверки необходимое условие положительности матрицы и одно достаточное условие положительности матрицы.

Пусть матрица $A > 0$, тогда для вектора $y = (0, \cdots, 0, y_i, 0, \cdots, 0)^T$, где $y_i \neq 0$, $(Ay, y) = a_{ii} y_i^2 > 0$ $(i = 1, \ldots, n)$. Отсюда следует, что у положительной матрицы A все диагональные элементы $a_{ii} > 0$. Эти же неравенства можно получить иначе. Пусть положительны все угловые миноры матрицы A. Следствием этого (см. [1]) является положительность всех главных миноров матрицы A. Так как элементы a_{ii}, находящиеся на главной диагонали матрицы A, являются главными минорами первого порядка, то для них выполнены неравенства $a_{ii} > 0$. Это и есть удобное для проверки необходимое условие положительности матрицы A.

Простым для проверки достаточным условием положительности симметричной матрицы $A = A^T$ является *условие диагонального преобладания*:

$$a_{ii} > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \; i = 1, 2, \ldots, n.$$

Покажем это. Пусть выполнено необходимое условие положительности матрицы A, то есть $a_{ii} > 0$, $i = 1, 2, \ldots, n$. Пусть выполнено условие диагонального преобладания, λ — любое из собственных чисел матрицы A и $\xi = (\xi_1, \cdots, \xi_n)^T$ — собственный вектор матрицы A, соответствующий этому собственному числу, то есть $A\xi = \lambda\xi$. Выберем максимальную по модулю компоненту собственного вектора ξ. Пусть $|\xi_i| = \max\limits_{1 \leq j \leq n} |\xi_j|$, тогда компонента с номером i векторного равенства $A\,\xi = \lambda\xi$ имеет вид

$$a_{ii}\xi_i + S = \lambda\xi_i, \text{ где } S = \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij}\xi_j.$$

Для $|S|$ справедлива следующая оценка:

$$|S| = \Big| \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij}\,\xi_j \Big| \leq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}||\xi_j| \leq |\xi_i| \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| < |\xi_i| a_{ii}.$$

Отсюда следует, что если $\xi_i > 0$, то $-a_{ii}\,\xi_i < S$. Тогда сумма $a_{ii}\,\xi_i + S = \lambda\xi_i > 0$ и, следовательно, $\lambda > 0$. Если $\xi_i < 0$, то $S < -a_{ii}\xi_i$, $a_{ii}\xi_i + S = \lambda\xi_i < 0$ и, как и в предыдущем случае, $\lambda > 0$.

Итак, выполнение условия диагонального преобладания гарантирует положительность всех собственных чисел матрицы A. Следовательно, симметричная матрица $A = A^T$ обладает свойством положительности (положительной определенности).

1.1.1 Метод квадратного корня (метод Холецкого)

Рассмотрим системы линейных алгебраических уравнений с симметричными и положительно определенными матрицами ($A^T = A > 0$). Для таких матриц возможно представление

$$A = LL^T, \tag{1.2}$$

где L — нижняя треугольная матрица. Такое представление матрицы A называют разложением Холецкого. Подставляя разложение (1.2) матрицы A в уравнение (1.1) и вводя обозначение $L^T y = \tilde{y}$, получим систему уравнений $L\tilde{y} = f$ относительно вектора $\tilde{y} = (\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_n)^T$. Поскольку матрица L нижняя треугольная, то решение этой системы осуществляется построчно. Из первого уравнения системы находится компонента $\tilde{y}_1$ вектора $\tilde{y}$. Из второго уравнения, зная $\tilde{y}_1$, находится компонента $\tilde{y}_2$ и так далее. Вычислив все компоненты вектора $\tilde{y}$, получаем для искомого вектора y систему уравнений $L^T y = \tilde{y}$ с верхней треугольной матрицей. Решение этой системы проводится построчно, начиная с последнего уравнения. Описанная процедура реализуется, если найдены элементы матрицы L.

Получим расчетные формулы для вычисления элементов матрицы L. Будем использовать обозначения $L = (l_{ij})$ и $L^T = (\bar{l}_{ij})$, где $\bar{l}_{ij} = l_{ji}$. Найдем элементы этих матриц, исходя из матричного равенства (1.2):

$$\Big(a_{ij}\Big) = \begin{pmatrix} l_{11} & 0 & \dots & 0 \\ l_{21} & l_{22} & & \vdots \\ \vdots & & \ddots & 0 \\ l_{n1} & l_{n2} & \dots & l_{nn} \end{pmatrix} \begin{pmatrix} \bar{l}_{11} & \bar{l}_{12} & \dots & \bar{l}_{1n} \\ 0 & \bar{l}_{22} & & \bar{l}_{2n} \\ \vdots & & \ddots & \vdots \\ 0 & \dots & 0 & \bar{l}_{nn} \end{pmatrix}.$$

Имеем

$$a_{ij} = \sum_{m=1}^{n} l_{im}\bar{l}_{mj} = \sum_{m=1}^{\min(i,j)} l_{im}\bar{l}_{mj} = \sum_{m=1}^{\min(i,j)} l_{im}l_{jm}. \tag{1.3}$$

Используем данное равенство для определения элементов l_{ij} в столбцах матрицы L.

При $j = 1$ соотношение (1.3) принимает вид $a_{i1} = l_{i1}l_{11}$, $i = 1, 2, \dots, n$. Отсюда $a_{11} = l_{11}^2$, где $a_{11} > 0$ в силу положительной определенности матри-

цы A. Поэтому справедлива формула $l_{11} = \sqrt{a_{11}}$. Определив l_{11}, вычислим остальные элементы первого столбца по формуле $l_{i1} = a_{i1}/l_{11}$, $i = 2, 3, \ldots, n$.

При $j = 2$ из того же соотношения (1.3) получим, что

$$a_{i2} = l_{i1}l_{21} + l_{i2}l_{22},\ i = 2, 3, \ldots, n. \tag{1.4}$$

Отсюда $l_{22}^2 = a_{22} - l_{21}^2$, где элемент $l_{21} = a_{21}/\sqrt{a_{11}}$ вычислен на предыдущем этапе. Извлекая корень, определим $l_{22} = \sqrt{(a_{11}a_{22} - a_{21}^2)/a_{11}}$. Операция извлечения корня корректна, так как

$$a_{11} > 0,\ a_{11}a_{22} - a_{12}^2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} > 0$$

в силу симметричности и положительной определенности матрицы A. Определив l_{22}, вычислим, используя (1.4), остальные элементы второго столбца по формуле $l_{i2} = (a_{i2} - l_{i1}l_{21})/l_{22}$, $i = 3, 4, \ldots, n$.

Далее, исходя из того, что элементы матрицы L в столбцах с номерами $1, 2, \ldots, m-1$ вычислены на предыдущих этапах, получим расчетные формулы для элементов столбца с номером m. При $j = m$ из (1.3) для $i = m, m+1, \ldots, n$ получим:

$$a_{im} = l_{i1}l_{m1} + l_{i2}l_{m2} + \ldots + l_{i(m-1)}l_{m(m-1)} + l_{im}l_{mm}. \tag{1.5}$$

Отсюда, полагая $i = m$, находим

$$l_{mm} = \sqrt{a_{mm} - l_{m1}^2 - \ldots - l_{m(m-1)}^2}.$$

Можно показать, что, как и ранее, под корнем получается отношение положительного углового минора $m-$го порядка матрицы A к положительному минору $(m-1)-$го порядка матрицы A. Определив l_{mm}, вычислим, используя (1.5), остальные элементы $m-$го столбца матрицы L

$$l_{im} = \frac{a_{im} - l_{i1}l_{m1} - \ldots - l_{i(m-1)}l_{m(m-1)}}{l_{mm}},\ i = m+1, m+2, \ldots, n.$$

Указанным способом последовательно определяются все элементы матрицы L. При этом подсчет числа арифметических действий, затрачиваемых на вычисление элементов матрицы L, дает величину $\approx n^3/6$, что в два раза меньше трудоемкости метода Гаусса.

1.1.2 Модифицированный метод квадратного корня

Рассмотрим уравнение $Ay = f$ при условии, что $\det A \neq 0$ и матрица A симметрична ($A^T = A$). Для таких матриц справедливо представление

$$A = LDL^T, \tag{1.6}$$

где L — нижняя треугольная матрица с единицами на главной диагонали, а D — диагональная матрица. Если указанное представление найдено, то решение исходной системы уравнений $Ay = f$ сводится к последовательному решению систем $L\hat{y} = f$, $D\tilde{y} = \hat{y}$ и $L^T y = \tilde{y}$ с нижней треугольной, диагональной и верхней треугольной матрицами, соответственно. Решение таких систем не представляет трудностей (см. пункт 1.1.1), и трудоемкость метода фактически сводится к нахождению матриц L и D.

Рассмотрим алгоритм нахождения элементов матриц L и D. Как и ранее, будем использовать обозначения $L = (l_{ij})$ и $L^T = (\bar{l}_{ij})$, где $\bar{l}_{ij} = l_{ji}$, а также $D = (d_{ij})$. Тогда равенство (1.6) примет вид

$$\Big(a_{ij}\Big) = \begin{pmatrix} 1 & 0 & \dots & 0 \\ l_{21} & 1 & & \vdots \\ \vdots & & \ddots & 0 \\ l_{n1} & l_{n2} & \dots & 1 \end{pmatrix} \begin{pmatrix} d_{11} & 0 & \dots & 0 \\ 0 & d_{22} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & 0 & \dots & d_{nn} \end{pmatrix} \begin{pmatrix} 1 & \bar{l}_{12} & \dots & \bar{l}_{1n} \\ 0 & 1 & & \bar{l}_{2n} \\ \vdots & & \ddots & \vdots \\ 0 & \dots & 0 & 1 \end{pmatrix}.$$

Отсюда, используя обозначение $(b_{ij}) = DL^T$, где $b_{ij} = d_{ii}\overline{l_{ij}} = d_{ii}l_{ji}$, получим

$$a_{ij} = \sum_{k=1}^{n} l_{ik}b_{kj} = \sum_{k=1}^{j} l_{ik}b_{kj} = \sum_{k=1}^{j} l_{ik}d_{kk}l_{jk}, \tag{1.7}$$

где $i \geq j$.

При $j = 1$ и $i \geq 1$ равенство (1.7) примет вид $a_{i1} = l_{i1}d_{11}$, так как $l_{11} = 1$. Отсюда находим $d_{11} = a_{11}$ и $l_{i1} = a_{i1}/d_{11}$, $i = 2, 3, \dots, n$. Формулы корректны при условии, что $a_{11} \neq 0$.

При $j = 2$ и $i \geq 2$ из (1.7) получим $a_{i2} = l_{i1}d_{11}l_{21} + l_{i2}d_{22}$, так как $l_{22} = 1$. Отсюда

$$d_{22} = a_{22} - l_{21}^2 d_{11} = a_{22} - \frac{a_{21}^2}{a_{11}} = \frac{a_{11}a_{22} - a_{12}a_{21}}{a_{11}} = \frac{\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}}{d_{11}}.$$

В правой части этого выражения находится отношения угловых миноров второго и первого порядка матрицы A. Определив d_{22}, вычислим элементы $l_{i2} = (a_{i2} - l_{i1}d_{11}l_{21})/d_{22}$, $i = 3, 4, \dots, n$. Формулы корректны при $d_{22} \neq 0$, то есть помимо углового минора первого порядка a_{11} должен быть отличен от нуля угловой минор второго порядка матрицы A.

Далее последовательно вычисляются элементы следующих столбцов матриц D и L. Приведем формулы для вычисления элементов столбца с номером

m:

$$d_{mm} = a_{mm} - \left(l_{m1}^2 d_{11} + \ldots + l_{mm-1}^2 d_{m-1m-1}\right) = \frac{\begin{vmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mm} \end{vmatrix}}{d_{11}\, d_{22} \cdots d_{m-1m-1}},$$
$$l_{im} = (a_{im} - l_{i1} d_{11} l_{m1} - \ldots - l_{im-1} d_{m-1m-1} l_{mm-1})/d_{mm},$$
$$i = m+1, m+2, \ldots, n.$$

Отметим, что эти формулы применимы в случае, когда угловые миноры всех порядков матрицы A отличны от нуля. То есть, условия применимости модифицированного метода квадратного корня совпадают с условиями применимости обычного метода Гаусса.

Подсчитав количество операций умножения и деления, необходимых для получения разложения (1.6) и для поиска вектора y, получим сложность данного метода, равную $\approx n^3/6$.

1.2 Итерационные методы решения систем линейных алгебраических уравнений

К итерационным методам относят те методы, в которых искомое решение системы линейных алгебраических уравнений строится как предел некоторой сходящейся последовательности. Эффективными итерационные методы оказываются при решении систем большой размерности (систем содержащих тысячи и более уравнений) с разреженными матрицами.

В итерационных методах, задав некоторый вектор y^0, называемый *начальным приближением*, строят по некоторому правилу последовательность векторов $y^1,\ y^2,\ \ldots,\ y^k,\ \ldots$. Верхний индекс k называют номером итерационного приближения. В общем случае правило вычисления итерационного приближения может зависть от номера k и очередное $(k+1)$-ое итерационное приближение может строиться по всем предыдущим итерационным приближениям, то есть

$$y^{k+1} = G_{k+1}(y^0,\ y^1,\ \ldots,\ y^k).$$

Выбор правила вычисления итерационных приближений определяет конкретный итерационный метод.

Определение. Итерационный метод называется $m-$*шаговым*, если каждое последующее итерационное приближение строится лишь по m предыдущим:

$$y^{k+1} = G_{k+1}(y^{k-m+1},\ \ldots,\ y^{k-1},\ y^k).$$

Рассмотрим одношаговые ($m = 1$) и двухшаговые ($m = 2$) итерационные методы. На примере одношаговых итерационных методов удобно обсуждать математический аппарат, используемый для исследования итерационных методов. В классе двухшаговых итерационных методов существуют достаточно эффективные, широко используемые на практике итерационные методы решения систем линейных алгебраических уравнений.

Определение. Если G_{k+1} — линейная функция своих аргументов, то такой итерационный метод называется *линейным*.

1.2.1 Линейные одношаговые итерационные методы

Согласно введенным определениям, любой линейный одношаговый итерационный метод имет вид:

$$y^{k+1} = S_{k+1}y^k + \psi_{k+1}, \tag{1.8}$$

где S_{k+1} — матрица, а ψ_{k+1} — вектор, задание которых определяет конкретный итерационный метод (размерность векторов и порядок матриц считаются одинаковыми).

Будем требовать от итерационного метода, чтобы вектор $y = A^{-1}f$ (искомое точное решение исходной задачи (1.1)) при подстановке вместо y^{k+1} и y^k обращал бы (1.8) в тождество:

$$A^{-1}f = S_{k+1}A^{-1}f + \psi_{k+1}.$$

Тогда вектор ψ_{k+1} можно представить в виде $\psi_{k+1} = Q_{k+1}f$. Здесь введено обозначение $Q_{k+1} = A^{-1} - S_{k+1}A^{-1}$. Формулу для Q_{k+1} домножим справа на матрицу A. Тогда для матрицы S_{k+1} получим представление $S_{k+1} = E - Q_{k+1}A$, где E — единичная матрица. Выражения для вектора ψ_{k+1} и матрицы S_{k+1} подставим в (1.8):

$$y^{k+1} = y^k - Q_{k+1}Ay^k + Q_{k+1}f.$$

Отсюда получим

$$(Q_{k+1})^{-1}\,\tau_{k+1}\frac{y^{k+1} - y^k}{\tau_{k+1}} + Ay^k = f,$$

где $\tau_{k+1} > 0$ — некоторое вещественное число. Вводя обозначение $B_{k+1} = (Q_{k+1})^{-1}\,\tau_{k+1}$, приходим к так называемой *канонической форме* записи одношагового линейного итерационного метода:

$$B_{k+1}\frac{y^{k+1} - y^k}{\tau_{k+1}} + Ay^k = f. \tag{1.9}$$

Конкретный линейный одношаговый метод определяется заданием матриц B_{k+1} и числовых параметров τ_{k+1}. В дальнейшем будем пользоваться следующей терминологией.

Определение. Если матрица $B_{k+1} = E$, то соответствующий итерационный метод называется *явным*, в противном случае — *неявным*.

В явных итерационных методах вектор y^{k+1} находится без решения вспомогательной системы линейных алгебраических уравнений с матрицей B_{k+1}.

Определение. Если $B_{k+1} = B$ и $\tau_{k+1} = \tau$, то метод называется *стационарным*, в противном случае — *нестационарным*.

Определение. Вектор $z^k = y^k - y$ (отклонение итерационного приближения y^k от точного решения y) будем называть *погрешностью итерационного приближения* на $k-$ой итерации.

Определение. Метод называется *сходящимся*, если $\|z^k\| \xrightarrow[k\to\infty]{} 0$ для некоторой выбранной нормы $\|\cdot\|$.

Точное решение y исходной системы ((1.1)) является пределом последовательности итерационных приближений и в большинстве случаев не может быть получено за конечное число арифметических действий. Пусть для начального итерационного приближения y^0, очередного итерационного приближения y^k и достаточно малой величиной $\varepsilon > 0$ выполняется неравенство

$$\|y^k - y\| \leqslant \varepsilon \|y^0 - y\|.$$

Неравенство означает, что на $k-$*ой* итерации погрешность итерационного приближения y^k не превышает погрешности начального приближения y^0, уменьшенной в $1/\varepsilon$ раз. Тогда итерационное приближение y^k будем считать приближенным решением, полученным с точностью ε. Для итерационных методов естественно ожидать, что, если неравенство выполняется для некоторого $k_0 = k_0(\varepsilon)$, то оно должно выполняться и для любого $k > k_0(\varepsilon)$. Число $k_0(\varepsilon)$ будем называть *минимальным числом итераций, необходимым для достижения заданной точности* ε. Чем меньше при прочих равных условиях $k_0(\varepsilon)$, тем эффективнее итерационный метод.

Следует отметить, что использовать указанное неравенство для контроля достигнутой точности на $k-$*ой* итерации и завершения итерационного процесса не представляется возможным, поскольку точное решение y неизвестно. Поэтому, для конкретных итерационных методов, используются априорные оценки для $k_0(\varepsilon)$.

1.2.2 Примеры одношаговых линейных итерационных методов

Запишем систему уравнений $Ay = f$ в виде:

$$\begin{cases} a_{11}y_1 + a_{12}y_2 + \ldots + a_{1n}y_n = f_1, \\ a_{21}y_1 + a_{22}y_2 + \ldots + a_{2n}y_n = f_2, \\ \ldots \\ a_{n1}y_1 + a_{n2}y_2 + \ldots + a_{nn}y_n = f_n. \end{cases} \tag{1.10}$$

Используя такую форму записи исходной системы линейных алгебраических уравнений, линейный одношаговый итерационный метод можно задать, расставляя итерационные индексы у компонент вектора y.

Метод Якоби.

Для построения итерационного метода Якоби припишем неизвестным y_i, имеющим в качестве сомножителей коэффициенты a_{ii}, $i = 1, 2, \ldots, n$, индекс следующего итерационного приближения $k+1$, а прочим неизвестным — индекс k:

$$\begin{cases} a_{11}y_1^{k+1} + a_{12}y_2^k \quad + \ldots + a_{1n}y_n^k \quad = f_1, \\ a_{21}y_1^k \quad + a_{22}y_2^{k+1} + \ldots + a_{2n}y_n^k \quad = f_2, \\ \ldots \\ a_{n1}y_1^k \quad + a_{n2}y_2^k \quad + \ldots + a_{nn}y_n^{k+1} = f_n. \end{cases}$$

Тогда, расчетная формула для вычисления компонент $k+1$ итерационного приближения вектора y примет вид:

$$y_i^{k+1} = \frac{1}{a_{ii}} \left(f_i - \sum_{j=1}^{i-1} a_{ij}y_j^k - \sum_{j=i+1}^{n} a_{ij}y_j^k \right), \; i = 1, 2, \ldots, n.$$

Запишем метод Якоби в канонической форме. Для это представим матрицу A в виде $A = L + D + R$, где $L = (l_{ij})$ — нижняя треугольная, $D = (d_{ij})$ — диагональная, $R = (r_{ij})$ — верхняя треугольная матрицы.

$$l_{ij} = \begin{cases} a_{ij}, & i > j, \\ 0, & i \leqslant j; \end{cases} \; d_{ij} = \begin{cases} a_{ij}, & i = j, \\ 0, & i \neq j; \end{cases} \; r_{ij} = \begin{cases} a_{ij}, & i < j, \\ 0, & i \geqslant j. \end{cases}$$

Тогда, система уравнений, определяющая метод Якоби, в матричной форме примет вид

$$Ly^k + Dy^{k+1} + Ry^k = f.$$

Прибавляя и вычитая в левой части Dy^k, получим

$$D(y^{k+1} - y^k) + Ay^k = f.$$

Отсюда следует, что в канонической форме записи (1.9) методу Якоби соответствует выбор $B_{k+1} = D$ и $\tau_{k+1} = 1$. Таким образом, метод Якоби является стационарным неявным одношаговым линейным итерационным методом с легко обратимой матрицей B_{k+1}.

Метод Зейделя.

Метод Зейделя получается, если в развернутой записи системы уравнений (1.10) приписать неизвестным y_i, имеющим в качестве сомножителей коэффициенты a_{ij} при $i \geqslant j$, индекс следующего итерационного приближения $k+1$, а прочим неизвестным — индекс k:

$$\begin{cases} a_{11}y_1^{k+1} + a_{12}y_2^k \quad + \ldots + a_{1n}y_n^k \quad = f_1, \\ a_{21}y_1^{k+1} + a_{22}y_2^{k+1} + \ldots + a_{2n}y_n^k \quad = f_2, \\ \ldots \\ a_{n1}y_1^{k+1} + a_{n2}y_2^{k+1} + \ldots + a_{nn}y_n^{k+1} = f_n. \end{cases}$$

Относительно y^{k+1} эта система решается следующим образом. Сначала из первого уравнения находится y_1^{k+1}, потом из второго — y_2^{k+1} и так далее. Расчетная формула для вычисления компонент $(k+1)$–го итерационного приближения вектора y имеет вид:

$$y_i^{k+1} = \frac{f_i - \sum\limits_{j=1}^{i-1} a_{ij}y_j^{k+1} - \sum\limits_{j=i+1}^{n} a_{ij}y_j^k}{a_{ii}}, \; i = 1, 2, \ldots, n.$$

Используя ранее введенное представление матрицы $A = L + D + R$, запишем метод Зейделя в матричной форме

$$Ly^{k+1} + Dy^{k+1} + Ry^k = f.$$

Добавляя и вычитая в левой части этого соотношения комбинацию $Ly^k + Dy^k$, получим:

$$(L + D)(y^{k+1} - y^k) + Ay^k = f.$$

Следовательно, в канонической форме записи (1.9) методу Зейделя соответствует выбор матрицы $B_{k+1} = L + D$ и итерационного параметра $\tau_{k+1} = 1$. То есть, метод Зейделя является стационарным неявным итерационным методом. Нижняя треугольная матрица $L + D$ легко обратима.

Метод релаксации.

Этот итерационный метод определяется выбором $B_{k+1} = D + \omega L$ и $\tau_{k+1} = \omega$ в канонической форме записи (1.9), где ω — заданный числовой параметр:

$$(D + \omega L)\frac{y^{k+1} - y^k}{\omega} + Ay^k = f.$$

Данный метод является стационарным и неявным с легко обратимой нижней треугольной матрицей B_{k+1}. Заметим, что рассмотренный выше метод Зейделя является частным случаем метода релаксации при $\omega = 1$. Выбирая параметр ω из диапазона $0 < \omega < 1$ получаем метод, который часто называют методом нижней релаксации. Выбирая $1 < \omega < 2$, получим метод верхней релаксации.

Метод простой итерации.

Этот метод является примером явного стационарного итерационного метода. В канонической форме записи (1.9) ему соответствует выбор $B_{k+1} = E$ и $\tau_{k+1} = \tau$:

$$\frac{y^{k+1} - y^k}{\tau} + Ay^k = f.$$

Метод Ричардсона.

Примером явного нестационарного итерационного метода является метод Ричардсона, который имеет вид:

$$\frac{y^{k+1} - y^k}{\tau_{k+1}} + Ay^k = f.$$

Здесь матрица $B_{k+1} = E$, а итерационные параметры τ_{k+1} рассчитываются на каждой итерации по некоторым формулам, к которым вернемся позже.

1.3 Условия сходимости одношаговых стационарных итерационных методов

1.3.1 Необходимые и достаточные условия сходимости

Для установления факта сходимости и проведения сравнительного анализа свойств различных итерационных методов необходим соответствующий

аналитический аппарат. В данном пункте рассмотрим некоторые элементы такого математического аппарата.

При выяснении возможной области применения конкретного стационарного одношагового итерационного метода может быть полезно следующее утверждение [3].

Теорема 1.1 (Достаточное условие сходимости стационарного одношагового итерационного метода)**.** *Пусть A — симметричная и положительно определенная матрица ($A^T = A > 0$), B — положительно определенная матрица ($B > 0$) и числовой параметр $\tau > 0$. Тогда итерационный процесс*

$$B\frac{y^{k+1} - y^k}{\tau} + Ay^k = f \tag{1.11}$$

сходится для любого начального приближения y^0, если выполнено матричное неравенство $B > \frac{\tau}{2}A$.

(Доказательство см. [3].)

Следствием из теоремы являются, например, два следующих утверждения, определяющие достаточные условия сходимости итерационного метода Якоби и метода релаксации. Напомним (см. пункт 1.2.2), что, если представить матрицу A в виде $A = L + D + R$, где L — нижняя треугольная, D — диагональная, R — верхняя треугольная матрицы, то методу Якоби в канонической форме записи (1.11) соответствует выбор $B = D$, $\tau = 1$, а методу релаксации $B = D + \omega L$, $\tau = \omega$.

Теорема 1.2. *Пусть $A^T = A$ и выполнено условие диагонального преобладания*

$$a_{ii} > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \ i = 1, 2, \ldots, n.$$

Тогда, при любом начальном приближении итерационный метод Якоби сходится.

▼ Доказательство. С учетом диагонального преобладания и симметричности матрицы A для $\forall\, y \neq 0$ выполнено:

$$0 < (Ay, y) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}y_iy_j \leqslant \sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}||y_iy_j| \leqslant \left\{2|y_iy_j| \leqslant y_i^2 + y_j^2\right\} \leqslant$$

$$\leqslant \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|y_i^2 + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|y_j^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|y_i^2 =$$

$$= \sum_{i=1}^{n} y_i^2 \Big(a_{ii} + \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \Big) < 2 \sum_{i=1}^{n} y_i^2 a_{ii} = 2(Dy, y).$$

Таким образом, показано, что $((2D - A)y, y) > 0 \ \forall \, y \neq 0$. Это условие эквивалентно матричному неравенству $D > \frac{1}{2}A$, которое в силу теоремы 1.1 достаточно для сходимости метода Якоби, поскольку $A^T = A > 0$ и $B = D > 0$.

▲ Утверждение доказано.

Теорема 1.3. *Пусть $A^T = A > 0$. Тогда метод релаксации с итерационным параметром $0 < \omega < 2$, является сходящимся итерационным методом для любого начального приближения.*

▼ Доказательство. Запишем достаточное условие сходимости $B > \frac{\tau}{2}A$ при $B = D + \omega L$, $\tau = \omega$ в виде

$$((2D + 2\omega L - \omega A)y, y) > 0 \quad (y \neq 0).$$

Так как $A^T = A$, то $A = L + D + L^T$ и, следовательно

$$(Ay, y) = (Ly.y) + (Dy, y) + (L^T y, y) = (Dy, y) + 2(Ly, y).$$

Тогда достаточное условие сходимости принимает вид:

$$2(Dy, y) + 2\omega(Ly, y) - 2\omega(Ly, y) - \omega(Dy, y) > 0.$$

В результате имеем неравенство $(2 - \omega)(Dy, y) > 0$, которое выполнено, так как в силу условий теоремы $(2 - \omega) > 0$ и $(Dy, y) > 0$.

▲ Утверждение доказано.

Сформулированные выше теоремы являются примерами достаточных условий сходимости. В случае, когда условия теорем не выполнены, сделать вывод о сходимости или расходимости соответствующих методов не представляется возможным. Полностью решить вопрос о сходимости конкретного итерационного метода можно лишь на основе какого-либо критерия его сходимости.

Рассмотрим пример критерия сходимости стационарных одношаговых итерационных методов (1.11). Предварительно получим уравнение для погрешности z^k. Подставляя в (1.11) представление итерационного приближения $y^k = z^k + y$, где y — точное решение исходного уравнения $Ay = f$, приходим к однородному уравнению $B\dfrac{z^{k+1} - z^k}{\tau} + Az^k = 0$, из которого следует, что

$z^{k+1} = (E - \tau B^{-1}A)z^k$. Введем обозначение $S = E - \tau B^{-1}A$. Матрица S называется *матрицей перехода*. Тогда уравнение для погрешности примет вид:

$$z^{k+1} = Sz^k. \tag{1.12}$$

Теорема 1.4 (Критерий сходимости одношагового стационарного итерационного метода)**.** *Итерационный метод (1.11) сходится для любого начального приближения y^0 тогда и только тогда, когда для всех собственных значений $\lambda(S)$ матрицы S выполнено неравенство $|\lambda(S)| < 1$.*

▼ Доказательство.

Предположим, что метод сходится при любом выборе начального приближения y^0. Пусть μ — собственный вектор матрицы S, отвечающий собственному значению λ. Рассмотрим вектор начального приближения $y^0 = \mu + y$, тогда $z^0 = \mu$. Из уравнения для погрешности (1.12) получим

$$z^k = Sz^{k-1} = S^2z^{k-2} = \ldots = S^kz^0 = S^k\mu = S^{k-1}\Big(S\mu\Big) = \lambda S^{k-1}\mu =$$
$$\ldots = \lambda^k\mu \;\Rightarrow \|z^k\| = |\lambda|^k\|\mu\|.$$

По предположению о сходимости $\|z^k\| = |\lambda|^k\|\mu\| \xrightarrow[k\to\infty]{} 0$. Поэтому приходим к неравенству $|\lambda| < 1$.

Доказательство достаточности проведем при дополнительном предположении, что матрица S имеет n линейно независимых собственных векторов μ_l, образующих базис n мерного пространства.

Пусть для любого собственного значения матрицы S выполнено неравенство $|\lambda| < 1$. Погрешность произвольного начального приближения z^0 представим в виде разложения по базису μ_l с коэффициентами c_l. Из уравнения (1.12) получим:

$$z^k = S^kz^0 = S^k\sum_{l=1}^{n} c_l\mu_l = \sum_{l=1}^{n} c_l\lambda_l^k\mu_l \;\Rightarrow \|z^k\| \leqslant \sum_{l=1}^{n} |c_l|\,|\lambda_l|^k\|\mu_l\| \leqslant \bar{\lambda}^kM.$$

Здесь $\bar{\lambda} = \max\limits_{1\leq l\leq n}|\lambda_l|$, $M = \sum\limits_{l=1}^{n}|c_l|\,\|\mu_l\|$. Так как $\bar{\lambda} < 1$, а $M = \text{const}$, то из последней оценки вытекает сходимость итерационного метода.

▲ Утверждение доказано.

Замечание. Дополнительное предположение существенно упрощает доказательство достаточности в предыдущей теореме. Однако доказательство можно провести (см. [3]) и для матрицы S общего вида на основе ее приведения к жордановой форме.

Наличие критерия сходимости казалось бы полностью решает вопрос об исследовании итерационного метода на сходимость. Однако применение рассмотренного критерия требует поиска всех собственных значений матрицы перехода. Данная задача может оказаться сложнее, чем решение исходной системы линейных алгебраических уравнений. Поэтому необходимы и другие, проще реализуемые на практике, способы исследования сходимости итерационных методов.

При практическом использовании итерационных методов важен не только факт сходимости, но и оценка количества итераций, необходимых для достижения заданной точности. Именно по этому показателю в совокупности с вычислительными затратами на осуществление одной итерации целесообразно оценивать качество метода.

Рассмотрим специальный класс удобных для исследования итерационных методов.

Определение. Итерационный метод *сходится со скоростью геометрической прогрессии* со знаменателем $\rho \in (0;1)$, если для погрешности итерационного приближения справедливо неравенство

$$||y^k - y|| \leqslant \rho^k ||y^0 - y||. \tag{1.13}$$

Пусть $\varepsilon > 0$ — некоторое достаточное малое число. Потребуем, чтобы погрешность итерационного приближения на итерации с номером k была бы не больше, чем погрешность начального итерационного приближения уменьшенная в $1/\varepsilon$ раз. Это означает, что $\|y^k - y\| \leqslant \varepsilon \|y^0 - y\|$. Для того, чтобы для итерационного метода, сходящегося со скоростью геометрической прогрессии, выполнялось это неравенство достаточно выполнения неравенства $\rho^k \leqslant \varepsilon$ верного при

$$k > k_0(\varepsilon) = \left[\frac{\ln(1/\varepsilon)}{\ln(1/\rho)}\right].$$

Определение. Число $k_0(\varepsilon)$ называется *минимальным числом итераций, необходимым для достижения заданной точности* ε.

Определение. Выражение $\ln(1/\rho)$ называется *скоростью сходимости* итерационного метода.

Чем больше скорость сходимости, тем меньше итераций необходимо выполнить для достижения требуемой точности вычисления итерационного приближения и тем лучше соответствующий итерационный метод.

1.3.2 Оценка скорости сходимости одношаговых стационарных методов

Рассмотрим пример утверждения позволяющего для конкретного итерационного метода, сходящегося со скоростью геометрической прогрессии, определить значение параметра ρ в неравенстве (1.13).

Далее будут использоваться следующие определения и утверждения (см. [3]).

1. Если $A^T = A$, то $A > 0 \Leftrightarrow \lambda > 0$.

2. Если $A^T = A > 0$, то $\exists\, A^{-1} > 0$.

3. Если $A^T = A, \rho > 0$, то $-\rho E < A < \rho E \Leftrightarrow A^2 < \rho^2 E$.

4. Если $A^T = A > 0$, то $\exists\, B \;:\; B^2 = A, B^T = B > 0$.

Определение. Матрица B называется *квадратным корнем* матрицы A и обозначается $A^{1/2}$ $(A^T = A \geqslant 0)$.

5. Если $A^T = A > 0, B^T = B > 0$, то $\alpha A > \beta B \Leftrightarrow \alpha B^{-1} > \beta A^{-1}$, где α, β — вещественные числа.

Замечание. В утверждениях 1,3,4,5 после слова «то» неравенства могут быть нестрогими.

Замечание. Аналогичные утверждения справедливы для линейных операторов в евклидовом пространстве H. При этом под C^T следует понимать оператор C^*.

Определение. *Матричной (операторной, энергетической) нормой* вектора v, порожденной симметричной положительно определенной матрицей A называется функционал $\|v\|_A = \sqrt{(Av, v)}$.

Замечание. $\|v\|_A = \sqrt{(A^{1/2}v, A^{1/2}v)} = \|A^{1/2}v\|$.

Ранее было показано, что погрешность z^k одношагового стационарного метода (1.11) удовлетворяет соотношению

$$z^{k+1} = Sz^k,\ S = E - \tau B^{-1}A.$$

Докажем следующие две леммы.

Лемма 1.1. *Пусть* $A^T = A > 0$, $B^T = B > 0$, $\rho > 0$ — *вещественное число, тогда неравенства*

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

необходимы и достаточны для того, чтобы при любых z^0 *для погрешности выполнялась оценка*

$$\|z^{k+1}\|_A \leqslant \rho\|z^k\|_A,\ k = 0, 1, \ldots\ .$$

▼ Доказательство. Обозначим $v^k = A^{1/2}z^k$, тогда

$$v^{k+1} = A^{1/2}z^{k+1} = A^{1/2}Sz^k = A^{1/2}SA^{-1/2}v^k = \tilde{S}v^k,$$

где $\tilde{S} = E - \tau C$, $C = A^{1/2}B^{-1}A^{1/2}$, $C^T = C > 0$. Имеем

$$\|z^{k+1}\|_A \leqslant \rho\|z^k\|_A \Leftrightarrow \|v^{k+1}\| \leqslant \rho\|v^k\| \Leftrightarrow (\tilde{S}^2v^k, v^k) \leqslant \rho^2(v^k, v^k) \Leftrightarrow$$
$$\Leftrightarrow \tilde{S}^2 \leqslant \rho^2 E \Leftrightarrow -\rho E \leqslant \tilde{S} \leqslant \rho E \Leftrightarrow \frac{1-\rho}{\tau}E \leqslant C \leqslant \frac{1+\rho}{\tau}E \Leftrightarrow$$
$$\Leftrightarrow \frac{1-\rho}{\tau}C^{-1} \leqslant E \leqslant \frac{1+\rho}{\tau}C^{-1} \Leftrightarrow$$
$$\Leftrightarrow \frac{1-\rho}{\tau}A^{-1/2}BA^{-1/2} \leqslant E \leqslant \frac{1+\rho}{\tau}A^{-1/2}BA^{-1/2} \Leftrightarrow$$
$$\Leftrightarrow \frac{1-\rho}{\tau}B \leqslant A^{1/2}EA^{1/2} \leqslant \frac{1+\rho}{\tau}B.$$

▲ Утверждение доказано.

Лемма 1.2. *При условиях леммы 1.1 справедлива оценка*

$$\|z^{k+1}\|_B \leqslant \rho\|z^k\|_B,\ k = 0, 1, \dots\ .$$

▼ Доказательство. Обозначив $v^k = B^{1/2}z^k$, получим, что $v^{k+1} = \tilde{S}v^k$, где $\tilde{S} = E - \tau C$, $C = B^{-1/2}AB^{-1/2}$. Далее аналогично предыдущему доказательству

$$\|z^{k+1}\|_B \leqslant \rho\|z^k\|_B \Leftrightarrow \frac{1-\rho}{\tau}E \leqslant C \leqslant \frac{1+\rho}{\tau}E \Leftrightarrow$$
$$\Leftrightarrow \frac{1-\rho}{\tau}B^{1/2}EB^{1/2} \leqslant A \leqslant \frac{1+\rho}{\tau}B^{1/2}EB^{1/2}.$$

▲ Утверждение доказано.

Теорема 1.5. *Пусть $A^T = A > 0$, $B^T = B > 0$, $\gamma_1 B \leqslant A \leqslant \gamma_2 B$, где γ_1, γ_2 — вещественные числа такие, что $\gamma_2 > \gamma_1 > 0$. Тогда при $\tau = 2/(\gamma_1 + \gamma_2)$ итерационный метод* (1.11) *сходится и для погрешности справедливы оценки*

$$\|z^k\|_A \leqslant \rho^k\|z^0\|_A,\ \|z^k\|_B \leqslant \rho^k\|z^0\|_B,\ k = 1, 2, \dots\ ,$$

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

▼ Доказательство.

В неравенстве $\gamma_1 B \leqslant A \leqslant \gamma_2 B$ константы γ_1 и γ_2 выразим через τ и ρ

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

Тогда неравенство примет вид

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

Согласно леммам 1.1, 1.2 неравенство равносильно оценкам погрешности

$$\|z^{k+1}\|_A \leqslant \rho \|z^k\|_A,\ \|z^{k+1}\|_B \leqslant \rho \|z^k\|_B,\ k = 0, 1, \dots\ .$$

То есть

$$\|z^k\|_A \leqslant \rho \|z^{k-1}\|_A \leqslant \dots \leqslant \rho^k \|z^0\|_A$$

и, аналогично,

$$\|z^k\|_B \leqslant \rho \|z^{k-1}\|_B \leqslant \dots \leqslant \rho^k \|z^0\|_B.$$

▲ Утверждение доказано.

Замечание. Скорость сходимости в случае, когда величина ξ мала, равна

$$\ln \frac{1}{\rho} = \ln \frac{1+\xi}{1-\xi} = \ln \left(1 + \frac{2\xi}{1-\xi}\right) \approx 2\xi = 2\frac{\gamma_1}{\gamma_2}$$

и, следовательно, число итераций, необходимое для достижения заданной точности ε равно

$$k_0(\varepsilon) = \frac{\ln(1/\varepsilon)}{\ln(1/\rho)} \approx \frac{\ln(1/\varepsilon)}{2\xi}.$$

Ускорить сходимость можно за счет увеличения константы γ_1 и уменьшения γ_2.

Замечание. Предполагая, что $A^T = A > 0$, $B^T = B > 0$, рассмотрим обобщенную задачу на собственные значения $A\mu = \lambda B\mu$, равносильную задаче поиска собственных значений и собственных векторов матрицы $B^{-1}A$. Уравнение задачи можно переписать в виде

$$(B^{-1/2}AB^{-1/2})(B^{1/2}\mu) = \lambda(B^{1/2}\mu) \ \Leftrightarrow\ C\tilde{\mu} = \lambda\tilde{\mu}.$$

Здесь $C = B^{-1/2}AB^{-1/2}$, $\tilde{\mu} = B^{1/2}\mu$. Поскольку $C^T = C > 0$, приходим к выводу, что все собственные значения λ матрицы C, они же собственные значения матрицы $B^{-1}A$, действительны и положительны. Тогда, неравенства

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

слева и справа умножая на $B^{-1/2}$, получаем

$$\gamma_1 E \leqslant B^{-1/2} A B^{-1/2} \leqslant \gamma_2 E \;\Leftrightarrow$$
$$\Leftrightarrow\; \gamma_1 \leqslant \lambda \leqslant \gamma_2,$$

где λ - любое собственное значение матрицы $B^{-1}A$. Отсюда вытекает, что $\gamma_1 = \lambda_{\min}(B^{-1}A)$, $\gamma_2 = \lambda_{\max}(B^{-1}A)$ — наиболее точные постоянные, с которыми выполняются неравенства $\gamma_1 B \leqslant A \leqslant \gamma_2 B$.

Определение. *Оптимальным итерационным параметром* метода (1.11) называется число

$$\tau = \frac{2}{\lambda_{\min}(B^{-1}A) + \lambda_{\max}(B^{-1}A)}.$$

Замечание. Оптимальный итерационный параметр минимизирует величину ρ на множестве всех положительных γ_1,γ_2, удовлетворяющих условиям $\gamma_1 B \leqslant A \leqslant \gamma_2 B$.

Замечание. Скорость сходимости максимальна, если выбрать $B = A$. Тогда $\rho = 0$ при $\gamma_1 = \lambda_{\min}(B^{-1}A) = \lambda_{\max}(B^{-1}A) = \gamma_2 = 1$, $\tau = 1$ и метод (1.11) дает точное решение уравнения $Ay = f$ на первой же итерации, поскольку $A(y^1 - y^0) + Ay^0 = f$. Однако для вычисления y^1 необходимо обратить матрицу A, что равносильно нахождению точного решения $y = A^{-1}f$.

Воспользуемся доказанной в этом пункте теоремой для сравнения скорости сходимости различных стационарных одношаговых итерационных методов. Тестировать итерационные методы будем на одной и тойже системе линейных алгебраических уравнений, которую назовем модельной задачей.

1.3.3 Модельная задача

Рассмотрим краевую задачу для ОДУ 2-го порядка:

$$-u''(x) = f(x),\; 0 < x < 1; u(0) = u(1) = 0.$$

Введем на отрезке $[0;1]$ равномерную разностную сетку с постоянным шагом h и узлами x_i

$$\Omega_h = \{x_i = ih;\; i = 0, 1, \ldots, N;\; hN = 1\}$$

и сопоставим дифференциальной задаче разностную схему

$$-y_{\bar{x}x,i} = f_i;\; i = 1, 2, \ldots, N-1;\; y_0 = y_N = 0. \tag{1.14}$$

Здесь $y_i = y(x_i)$, $f_i = f(x_i)$ — сеточные функции, определенные в узлах сетки x_i, а $y_{\bar{x}x,i}$ — сокращенная запись разностного отношения

$$y_{\bar{x}x,i} = \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2}.$$

Пусть функция $u(x)$ — достаточно гладкое решение дифференциальной задачи. Подставляя ее значения в узлах сетки в разностную схему, получим, используя разложения по формуле Тейлора с центром в узле x_i, что сеточная функция

$$\psi_i = \frac{u(x_{i-1}) - 2u(x_i) + u(x_{i+1})}{h^2} + f(x_i) =$$
$$= u''(x_i) + \frac{h^2}{12}u^{(4)}(x_i) + O(h^4) + f(x_i) = \frac{h^2}{12}u^{(4)}(x_i) + O(h^4) = O(h^2).$$

Функция ψ_i называется погрешностью аппроксимации разностной схемы на решении дифференциальной задачи, а равенство $\psi_i = O(h^2)$ означает, что разностная схема аппроксимирует исходную задачу со вторым порядком по параметру h.

Введем векторы

$$y = (y_1, y_2, \ldots, y_{N-1})^T, \quad f = (f_1, f_2, \ldots, f_{N-1})^T;$$

и матрицу

$$A = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & \ldots & 0 & 0 & 0 \\ -1 & 2 & -1 & \ldots & 0 & 0 & 0 \\ 0 & -1 & 2 & \ldots & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & -1 & 2 & -1 \\ 0 & 0 & 0 & \ldots & 0 & -1 & 2 \end{pmatrix}.$$

Тогда в матричной форме разностная схема (1.14) может быть записана в виде системы линейных алгебраических уравнений $Ay = f$, где матрица A является симметричной матрицей ($A^T = A$).

Именно эту систему линейных алгебраических уравнений в дальнейшем будем использовать как модельную задачу для тестирования и сопоставления между собой различных итерационных методов.

Для доказательства положительной определенности матрицы A, найдем ее собственные значения и убедимся в их положительности. Для этого рассмотрим разностную задачу на собственные значения

$$\mu_{\bar{x}x,i} + \lambda\mu_i = 0,\ \mu_0 = \mu_N = 0,\ i = 1, 2, \ldots, N-1,\ hN = 1. \tag{1.15}$$

Нахождение чисел λ_l и соответствующих им сеточных функций μ_i^l ($l = 1, 2, \ldots, N-1,\ i = 0, 1, \ldots, N$), являющихся решением этой задачи, эквивалентно нахождению собственных значений и собственных векторов матрицы A.

Лемма 1.3. *Решения задачи* (1.15) *имеют вид:*

$$\lambda_l = \frac{4}{h^2}\sin^2\frac{\pi l}{2N},\ \mu_i^l = \sin\frac{\pi l i}{N},\ l = 1, 2, \ldots, N-1,\ i = 0, 1, \ldots, N.$$

▼ Доказательство. По аналогии с соответствующей дифференциальной задачей на собственные значения для дифференциального оператора второй производной будем искать решение в виде

$$\mu_i = \sin(\alpha i),\ i = 0, 1, \ldots, N.$$

Подставляя $\mu_i = \sin(\alpha i)$ в уравнение (1.15)

$$\mu_{i-1} - 2(1 - \lambda h^2/2)\mu_i + \mu_{i+1} = 0,\ i = 1, 2, \ldots, N-1,$$

получим $2\sin(\alpha i)\cos\alpha - 2(1 - \lambda h^2/2)\sin(\alpha i) = 0$. Из этого следует

$$\cos\alpha = 1 - \frac{\lambda h^2}{2},\ \lambda = \frac{4}{h^2}\sin^2\frac{\alpha}{2}.$$

Поскольку $\mu_0 = \sin(\alpha 0) = 0$, осталось учесть граничное условие $\mu_N = 0$:

$$\sin(\alpha N) = 0 \Rightarrow \alpha = \pi l/N,\ l = 1, 2, \ldots, N-1.$$

▲ Утверждение доказано.

Тем самым показано, что собственные значения λ_l, $l = 1, 2, \ldots, N-1$ матрицы A различны и положительны. Таким образом, $A^T = A > 0$. При этом с учетом равенства $hN = 1$

$$\lambda_{\min} = \lambda_1 = \frac{4}{h^2}\sin^2\frac{\pi h}{2}, \quad \lambda_{\max} = \lambda_{N-1} = \frac{4}{h^2}\cos^2\frac{\pi h}{2}.$$

Используем построенную модельную задачу для тестирования итерационных методов. Применим для решения системы линейных алгебраических уравнений (1.14) с симметричной и положительно определенной матрицей итерационный метод простой итерации с оптимальным итерационным параметром.

Определение. Методом *простой итерации* для решения системы $Ay = f$ называется итерационный метод (1.11) при $B = E$, то есть

$$\frac{y^{k+1} - y^k}{\tau} + Ay^k = f;\ k = 0, 1, \ldots\ .$$

Тогда справедливо следствие.

Следствие. *Пусть* $A^T = A > 0$, $\lambda_{\min}(A)$ *и* $\lambda_{\max}(A)$ *— соответственно минимальное и максимальное собственные значения матрицы* A. *Тогда для метода простой итерации при*

$$\tau = \frac{2}{\lambda_{\min}(A) + \lambda_{\max}(A)}$$

справедлива оценка погрешности

$$\|z^k\| \leqslant \rho^k \|z^0\|,\ k = 1, 2, \ldots,\ \text{где } \rho = \frac{1-\xi}{1+\xi},\ \xi = \frac{\lambda_{\min}(A)}{\lambda_{\max}(A)}.$$

Замечание. Можно показать, что метод Якоби для решения уравнений разностной схемы (1.14) совпадает с методом простой итерации с оптимальным итерационным параметром (см. пункт 1.3.2) для модельной задачи.

Получим асимптотические оценки при $N \to \infty$ для скорости сходимости $\ln(1/\rho)$ и минимального числа итераций $k_0(\varepsilon)$, необходимых для достижения заданной точности ε:

$$\ln\frac{1}{\rho} = \ln\left(1 + \frac{2\xi}{1-\xi}\right) \approx 2\xi \ = 2\,\mathrm{tg}^2\,\frac{\pi h}{2} \approx \frac{\pi^2 h^2}{2} = \frac{\pi^2}{2N^2},$$
$$k_0(\varepsilon) = \frac{\ln(1/\varepsilon)}{\ln(1/\rho)} \approx \frac{2N^2}{\pi^2}\ln\frac{1}{\varepsilon}.$$

Пусть, например, $\varepsilon = 0,5 \cdot 10^{-4}$, тогда $\ln(1/\varepsilon) \approx 9,9$ и $k_0(\varepsilon) \approx 2N^2$. Это означает, что для нахождения приближенного решения разностных уравнений (1.14) с заданной точностью ε на сетке с $N = 10$ узлами требуется выполнить порядка 200 итераций, а для $N = 100$ — уже 20000 итераций. Такой быстрый рост числа итераций при увеличении размерности модельной задачи является характерной особенностью метода простой итерации с оптимальным итерационным параметром и метода Якоби.

1.4 Попеременно–треугольный итерационный метод

1.4.1 Алгебраическая теория

В пункте 1.2.2 некоторые стандартные итерационные методы приводились к каноническому виду

$$B\frac{y^{k+1} - y^k}{\tau} + Ay^k = f.$$

Здесь же проиллюстрируем возможность построения итерационного метода путем специального выбора матрицы B в канонической форме записи итерационного процесса.

Далее будем предполагать, что матрица системы уравнений $Ay = f$ симметрична и положительно определена. Введем матрицу

$$R = (r_{ij}), \quad r_{ij} = \begin{cases} a_{ij}, & i > j; \\ 0{,}5\, a_{ij}, & i = j; \\ 0, & i < j; \end{cases} \quad i, j = 1, 2, \ldots, n.$$

Матрица R является нижней треугольной матрицей, а транспонированная по отношению к ней матрица R^T — верхней треугольной. Матрица A представима в виде $A = R + R^T$, причем

$$0 < (Av, v) = ((R + R^T)v, v) = 2(Rv, v), \ \forall\, v \neq 0 \Rightarrow R, R^T > 0.$$

Для попеременно–треугольного итерационного метода матрица B определяется как произведение

$$B = (E + \omega R^T)(E + \omega R),$$

где E — единичная матрица, а $\omega > 0$ — числовой параметр. Такой выбор матрицы B обусловлен следующими обстоятельствами.

1) Используя вспомогательное промежуточное значение $y^{k+1/2}$, где

$$(E + \omega R^T) \underbrace{(E + \omega R)y^{k+1}}_{y^{k+1/2}} = \underbrace{(B - \tau A)y^k + \tau f}_{\varphi_k},$$

решение на новой итерации легко находится в два этапа:

$(E + \omega R^T)y^{k+1/2} = \varphi_k$ — система с верхней треугольной матрицей;
$(E + \omega R)y^{k+1} = y^{k+1/2}$ — система с нижней треугольной матрицей.

Замечание. Отсюда название метода.

2) Поскольку $B^T = B > 0$, так как

$$B = E + \omega A + \omega^2 R^T R \ \Rightarrow B^T = B,$$
$$(Bv, v) = ((E + \omega R)v, (E + \omega R)v) > 0 \ \Rightarrow B > 0,$$

то для попеременно-треугольного итерационного метода можно использовать полученные ранее оценки сходимости.

Лемма 1.4. *Пусть существуют положительные постоянные δ и Δ такие, что выполнены матричные неравенства $A \geqslant \delta E$, $4R^TR \leqslant \Delta A$. Тогда для матриц $A = R + R^T$ и $B(\omega) = (E + \omega R^T)(E + \omega R)$ справедливы неравенства*

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B, \text{ где } \gamma_1 = \left(\frac{1}{\delta} + \omega + \frac{\omega^2\Delta}{4}\right)^{-1}, \ \gamma_2 = \frac{1}{2\omega}.$$

▼ Доказательство.

$$B(\omega) = E + \omega A + \omega^2 R^T R \leqslant \left(\frac{1}{\delta} + \omega + \frac{\omega^2\Delta}{4}\right) A;$$
$$B(\omega) = E + \omega A + \omega^2 R^T R = E - \omega A + \omega^2 R^T R + 2\omega A =$$
$$= (E - \omega R^T)(E - \omega R) + 2\omega A \Rightarrow B(\omega) \geqslant 2\omega A.$$

▲ Утверждение доказано.

Замечание. Тем самым показано, что нахождение постоянных γ_1 и γ_2 сводится к нахождению постоянных δ и Δ. При выполнении неравенств $A \geqslant \delta E$, $4R^TR \leqslant \Delta A$ для произвольного вектора $v \neq 0$ имеем

$$\delta\|v\|^2 \leqslant (Av, v) = \frac{(Av, v)^2}{(Av, v)} = \frac{4(Rv, v)^2}{(Av, v)} \leqslant \frac{4\|Rv\|^2\|v\|^2}{(Av, v)} =$$
$$= \frac{4(R^TRv, v)\|v\|^2}{(Av, v)} \leqslant \frac{\Delta(Av, v)\|v\|^2}{(Av, v)} = \Delta\|v\|^2.$$

Отсюда следует, что $\delta \leqslant \Delta$. В качестве константы δ можно взять минимальное собственное значение $\lambda_{\min}(A)$ матрицы A. Также отметим, что, поскольку $(Av, v) \leqslant \Delta\|v\|^2$, то выполняется неравенство $\Delta \geqslant \lambda_{\max}(A)$, где $\lambda_{\max}(A)$ — максимальное собственное значение матрицы A.

Теорема 1.6. *Предположим, что для симметричной и положительно определенной матрицы $A = R + R^T$ известны положительные постоянные δ и Δ, при которых выполнены неравенства $A \geqslant \delta E$, $4R^TR \leqslant \Delta A$. Пусть*

$$\omega = \frac{2}{\sqrt{\delta\Delta}}, \ \tau = \frac{2}{\gamma_1 + \gamma_2}, \text{ где } \gamma_1 = \frac{\delta}{2(1 + \sqrt{\eta})}, \ \gamma_2 = \frac{\delta}{4\sqrt{\eta}}, \ \eta = \frac{\delta}{\Delta}.$$

Тогда попеременно–треугольный итерационный метод сходится и для его погрешности справедлива оценка

$$\|y^k - y\|_A \leqslant \rho^k\|y^0 - y\|_A, \text{ где } \rho = \frac{1 - \sqrt{\eta}}{1 + 3\sqrt{\eta}}.$$

▼ Доказательство. Согласно теореме 1.5 для выполнения требуемой оценки погрешности с константой

$$\rho(\omega) = \frac{1-\xi}{1+\xi} = 1 - \frac{2\xi}{1+\xi} = 1 - \frac{2}{1+\xi^{-1}},\ \xi(\omega) = \frac{\gamma_1(\omega)}{\gamma_2(\omega)}$$

достаточно положить $\tau = 2/(\gamma_1+\gamma_2)$. Выберем параметр $\omega > 0$ так, чтобы минимизировать $\rho(\omega)$. Для этого достаточно найти значение $\omega = \omega_0$, при котором функция $\xi^{-1}(\omega)$ достигает минимума. Согласно лемме 1.4

$$\xi^{-1}(\omega) = \frac{\gamma_2(\omega)}{\gamma_1(\omega)} = \frac{1}{2\omega}\left(\frac{1}{\delta} + \omega + \frac{\omega^2\Delta}{4}\right) = \frac{1}{2} + \frac{1}{2}\left(\frac{1}{\omega\delta} + \frac{\omega\Delta}{4}\right) =$$
$$= \frac{1}{2} + \frac{1}{2}\left(\frac{1}{\sqrt{\omega\delta}} - \frac{\sqrt{\omega\Delta}}{2}\right)^2 + \frac{1}{2}\sqrt{\frac{\Delta}{\delta}}.$$

Отсюда находим точку минимума $\omega_0 = 2/\sqrt{\delta\Delta}$. Подставляя это значение ω_0 в выражения для γ_1 и γ_2, получим

$$\gamma_1(\omega_0) = \left(\frac{1}{\delta} + \frac{2}{\sqrt{\delta\Delta}} + \frac{1}{\delta}\right)^{-1} = \frac{1}{2}\left(\frac{\sqrt{\delta}+\sqrt{\Delta}}{\delta\sqrt{\Delta}}\right)^{-1} = \frac{\delta}{2(1+\sqrt{\eta})},$$
$$\gamma_2(\omega_0) = \frac{\sqrt{\delta\Delta}}{4} = \frac{\delta}{4\sqrt{\eta}},\ \text{где } \eta = \frac{\delta}{\Delta} \in (0;1].$$
$$\text{Тогда}\xi(\omega_0) = \frac{\gamma_1}{\gamma_2} = \frac{2\sqrt{\eta}}{1+\sqrt{\eta}},\ \rho(\omega_0) = \frac{1-\xi}{1+\xi} = \frac{1-\sqrt{\eta}}{1+3\sqrt{\eta}} \in [0;1).$$

▲ Утверждение доказано.

Применение попеременно–треугольного метода к модельной задаче.

Обсудим применение попеременно–треугольного итерационного метода к модельной задаче (1.14). Напомним, что для модельной задачи $A^T = A > 0$.

Для того, чтобы применить попеременно–треугольный итерационный метод необходимо знать постоянные δ и Δ, определяющие параметры метода. Как уже отмечалось, в качестве δ и Δ можно взять минимальное собственное значение матрицы A и, соответственно, максимальное собственное значение матрицы A, то есть

$$\delta = \lambda_{\min}(A) = \frac{4}{h^2}\sin^2\frac{\pi h}{2},\ \Delta = \lambda_{\max}(A) = \frac{4}{h^2}\cos^2\frac{\pi h}{2}.$$

Согласно теореме 1.6 при выборе параметров метода

$$\omega = \frac{2}{\sqrt{\delta\Delta}},\ \tau = \frac{2}{\gamma_1 + \gamma_2},\ \text{где}$$
$$\gamma_1 = \frac{\delta}{2(1+\sqrt{\eta})},\ \gamma_2 = \frac{\delta}{4\sqrt{\eta}},\ \eta = \frac{\delta}{\Delta} = \operatorname{tg}^2\frac{\pi h}{2},$$

справедлива оценка погрешности с постоянной

$$\rho = \frac{1-\sqrt{\eta}}{1+3\sqrt{\eta}} = \frac{1-\operatorname{tg}\frac{\pi h}{2}}{1+3\operatorname{tg}\frac{\pi h}{2}} \approx \left(1-\frac{\pi h}{2}\right)\left(1-3\frac{\pi h}{2}\right) \approx 1-2\pi h.$$

Отсюда минимальное число итераций, необходимое для достижения заданной точности ε, равно

$$k_0(\varepsilon) = \frac{\ln(1/\varepsilon)}{\ln(1/\rho)} \approx \frac{\ln(1/\varepsilon)}{-\ln(1-2\pi h)} \approx \frac{\ln(1/\varepsilon)}{2\pi h} = \frac{N}{2\pi}\ln\frac{1}{\varepsilon}.$$

При $\varepsilon = 0,5\cdot 10^{-4}$ получаем, что $k_0(\varepsilon) \approx 1,6N$.

Замечание. Напомним, что для метода простой итерации число итераций при больших N оценивалось как $O(N^2)$. То есть попеременно–треугольный итерационный метод обеспечивает на порядок более быструю сходимость.

1.4.2 Чебышевский набор итерационных параметров

Используем для решения уравнения $Ay = f$ следующую нестационарную итерационную схему

$$B\frac{y^l - y^{l-1}}{\tau_l} + Ay^{l-1} = f, l = 1, 2, \ldots, k.$$

Повысить скорость сходимости итерационного метода можно за счет использования переменного итерационного параметра τ_l, зависящего от номера итерации. При фиксированном числе итераций k можно указать набор итерационных параметров τ_1, τ_2, …, τ_k , обеспечивающий наилучшую скорость сходимости вне зависимости от выбора начального приближения.

Перейдем от y^l-го итерационного приближения к погрешности $z^l = y^l - y$. Тогда получим следующее равенство

$$B\frac{z^l - z^{l-1}}{\tau_l} + Az^{l-1} = 0.$$

Отсюда выразим погрешность на l-й итерации z^l

$$z^l = (E - \tau_l B^{-1}A)z^{l-1}, \quad l = \overline{1,\ k}.$$

Рекурсивно применяя эту формулу для z^k, получим выражение для z^k

$$z^k = (E - \tau_k B^{-1}A)(E - \tau_{k-1}B^{-1}A)\dots(E - \tau_1 B^{-1}A)z^0.$$

Фиксируем число итераций (k) и постараемся выбрать τ_l так, чтобы погрешность на k-й итерации была бы минимально возможной:

$$||z^k|| \longrightarrow \inf_{\tau_l}.$$

Эта задача имеет точное решение (см. [2]). Соответствующий набор итерационных параметров τ_l принято называть чебышевским набором итерационных параметров.

Справедлива следующая теорема (доказательство см. в [2]).

Теорема 1.7. *Пусть* $A^T = A > 0$, $B^T = B > 0$, *а* τ_l *вычисляются по формуле:*

$$\tau_l = \frac{\tau_0}{1 + \rho_0 t_l},\ \textit{где}\ \ \tau_0 = \frac{2}{\lambda_{min}(B^{-1}A) + \lambda_{max}(B^{-1}A)},\quad \rho_0 = \frac{1-\xi}{1+\xi},$$

$$\xi = \frac{\lambda_{min}(B^{-1}A)}{\lambda_{max}(B^{-1}A)},\quad t_l = \cos\frac{(2l-1)\pi}{2k},\ l = \overline{1,\,k}.$$

Тогда погрешность $||y^k - y||_A$ *будет минимально возможной, и для нее справедлива оценка*

$$||y^k - y||_A \leqslant q_k ||y^0 - y||_A,$$

где $q_k = \dfrac{\rho_1^k}{1 + \rho_1^{2k}},\ \rho_1 = \dfrac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}.$

Замечание. Выясним, при каких значениях k выполняется условие выхода из итерационного процесса $||y^k - y||_A \leqslant \varepsilon ||y^0 - y||_A$.

Из предыдущей теоремы следует, что это условие будет выполнено, если $q_k \leqslant \varepsilon$. То есть,

$$\frac{\rho_1^k}{1 + \rho_1^{2k}} \leqslant \varepsilon \iff \varepsilon(\rho_1^k)^2 - \rho_1^k + \varepsilon \geqslant 0.$$

Корнями этого квадратного неравенства будут

$$\rho_1^k = \frac{1 \pm \sqrt{1 - 4\varepsilon^2}}{2\varepsilon} \approx \frac{1 \pm (1 - 2\varepsilon^2)}{2\varepsilon} = \left[\begin{array}{l} \rho_1^k = \varepsilon; \\ \rho_1^k = \dfrac{1}{\varepsilon} - \varepsilon. \end{array}\right.$$

Отсюда следует, что $q_k \leqslant \varepsilon$ при

$$\left[\begin{array}{l} \rho_1^k \leqslant \varepsilon; \\ \rho_1^k \geqslant \dfrac{1}{\varepsilon} - \varepsilon. \end{array}\right.$$

Из условий предыдущей теоремы следует, что ρ_1^k меньше единицы. Поэтому

$$q_k \leqslant \varepsilon \iff \rho_1^k \leqslant \varepsilon.$$

Таким образом, условие выхода из итерационного процесса выполнено при

$$k \geqslant k_0(\varepsilon) = \left[\frac{\ln\frac{1}{\varepsilon}}{\ln\frac{1}{\rho_1}}\right].$$

Для скорости сходимости $(\ln\frac{1}{\rho_1})$ верна оценка

$$\ln\frac{1}{\rho_1} = \ln\frac{1+\sqrt{\xi}}{1-\sqrt{\xi}} \approx \ln(1+2\sqrt{(\xi)}) \approx 2\sqrt{\xi}.$$

Применение метода Ричардсона к модельной задаче.

Воспользуемся сформулированной теоремой для оценки числа итераций $k_0(\varepsilon)$ в случае применения метода Ричардсона ($B = E$) с чебышевскими итерационными параметрами для решения модельной задачи.

В рассматриваемом случае собственные значения $\lambda_{min}(B^{-1}A)$ и $\lambda_{max}(B^{-1}A)$ равны

$$\lambda_{min}(B^{-1}A) = \lambda_{min}(A) = \frac{4}{h^2}\sin^2(\frac{\pi h}{2}), \quad \lambda_{max}(B^{-1}A) = \lambda_{max}(A) = \frac{4}{h^2}\cos^2(\frac{\pi h}{2}),$$

поэтому $\xi = \frac{\lambda_{min}}{\lambda_{max}} = \mathrm{tg}^2(\frac{\pi h}{2})$.

Пусть, как и прежде, $\varepsilon = 0,5 \cdot 10^{-4} \approx e^{-10}$. В этом случае, в соответствии с замечанием к теореме, получаем, что

$$k_0(\varepsilon) = \frac{\ln\frac{1}{\varepsilon}}{\ln\frac{1}{\rho_1}} \approx \frac{10}{2\,\mathrm{tg}(\frac{\pi h}{2})} \approx \frac{10}{\pi h} = \ \{\mathsf{h} = \frac{1}{\mathsf{N}}\} \ = \frac{10}{\pi}N \approx 3,2N.$$

Тогда при $N = 10$ нам понадобится 32 итерации, а в случае $N = 100 - 320$ итераций, что сопоставимо с соответствующими величинами для попеременно–треугольного итерационного метода.

Упорядоченный набор чебышевских параметров.

Замечание. Оценка числа итераций $k_0(\varepsilon)$ не зависит от порядка, в котором применяются итерационные параметры τ_l, однако этот порядок существенно влияет на вычислительную устойчивость алгоритма. При практическом применении данного метода используется специальный алгоритм построения упорядоченного набора итерационных параметров, обеспечивающий устойчивость вычислений.

Описанный в предыдущем разделе итерационный процесс гарантирует минимальное значение нормы погрешности итерационного приближения на k-й итерации. Расчет k-ого итерационного приближения осуществляется последовательно от y^0 до y^k в соответствии с используемой расчетной схемой итерационного процесса. Данная расчетная схема не гарантирует монотонного по итерациям убывания нормы погрешности итерационного приближения. Поэтому, при реализации итерационного метода с чебышевским набором параметров, возможен рост нормы погрешности на нескольких соседних итерациях, что может приводить к возникновению неприятных ситуаций (переполнению арифметических устройств). Для предотвращения таких неприятностей рекомендуется использовать, так называемый, упорядоченный чебышевский набор итерационных параметров.

В рассматриваемом методе формула для итерационного параметра τ_l содержит параметр $t_l = \cos(\frac{(2l-1)\pi}{2k})$. Запишем это выражение следующим образом

$$t_l = \cos\left(\frac{(2l-1)\pi}{2k}\right) = \cos\left(\frac{\pi}{2k}\,\theta_l^k\right), \quad l = \overline{1,\,k},$$

где θ_l^k нечетное число из множества θ^k, состоящего из нечетных чисел от 1 до $2k-1$.

Сформулируем, как пример, правило упорядочевания элементов множества θ^k в одном частном случае. Для произвольного k правило упорядочевания элементов множества θ^k можно найти в [2].

Пусть $k = 2^p$. Тогда элементы θ_l^k будем упорядочивать используя следующие рекуррентные формулы

$$\theta_1^1 = 1, \quad \theta_{2i-1}^{2m} = \theta_i^m, \quad \theta_{2i}^{2m} = 4m - \theta_{2i-1}^{2m}, \quad i = \overline{1,\,m}, \quad m = 1, 2, 4, \ldots, 2^{p-1}.$$

Проиллюстрируем эти формулы на примере $k = 2^3 = 8$. Так как $p = 3$, то $m = 1, 2, 4$. Итак

$$\begin{aligned}
& \theta^1 = \left\{\theta_1^1\right\} = \{1\} \\
m = 1: \quad & \theta^2 = \left\{\theta_1^2, \theta_2^2\right\} = \left\{\theta_1^2 = \theta_1^1 = 1, \theta_2^2 = 4 - \theta_1^1 = 3\right\} \\
m = 2: \quad & \theta^4 = \left\{\theta_1^4, \theta_2^4, \theta_3^4, \theta_4^4\right\} = \left\{\theta_1^4 = \theta_1^2 = 1, \theta_2^4 = 8 - \theta_1^4 = 7,\right.
\end{aligned}$$

$$\theta_3^4 = \theta_2^2 = 3, \theta_4^4 = 8 - \theta_3^4 = 5\}$$
$$m = 4: \quad \theta^8 = \{\theta_1^8, \theta_2^8, \ldots, \theta_8^8\} = \{\theta_1^8 = \theta_1^4 = 1, \theta_2^8 = 16 - \theta_1^8 = 15,$$
$$\theta_3^8 = \theta_2^4 = 7, \theta_4^8 = 16 - \theta_3^8 = 9, \theta_5^8 = \theta_3^4 = 3,$$
$$\theta_6^8 = 16 - \theta_5^8 = 13, \theta_7^8 = \theta_4^4 = 5, \theta_8^8 = 16 - \theta_7^8 = 11\}.$$

1.4.3 Попеременно–треугольный итерационный метод с упорядоченным набором чебышевских параметров

Рассмотрим итерационный метод соединяющий в себе достоинства как попеременно–треугольного метода, так и метода с чебышевским набором итерационных параметров. Пусть система линейных алгебраических уравнений $Ay = f$ имеет симметричную и положительно определенную матрицу A. Используем для ее решения линейный одношаговый нестационарный неявный итерационный метод, для которого справедливо следующее утверждение.

Теорема 1.8. *Пусть для решения уравнения $Ay = f$ используется итерационная схема*

$$B\frac{y^l - y^{l-1}}{\tau_l} + Ay^{l-1} = f, \quad l = \overline{1,\, k},$$

где $B = (E+\omega R^T)(E+\omega R), R+R^T = A$. Пусть известны положительные постоянные δ и Δ такие, что выполнены матричные неравенства

$$A \geqslant \delta E, \ \Delta A \geqslant 4R^T R.$$

Пусть

$$\tau_l = \frac{\tau_0}{1 + \rho_0 t_l}, \quad \tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad \rho_0 = \frac{\gamma_2 - \gamma_1}{\gamma_2 + \gamma_1},$$

$$\gamma_1 = \frac{\delta}{2(1 + \sqrt{\eta})}, \quad \gamma_2 = \frac{\delta}{4\sqrt{\eta}}, \quad \eta = \frac{\delta}{\Delta},$$

$$t_l = \cos\left(\frac{\pi}{2k}\,\theta_l^k\right), \quad \omega = \frac{2}{\sqrt{\delta\Delta}}.$$

Тогда, для выполнения на k-ой итерации неравенства

$$||A(y^k - y)||_{B^{-1}} \leqslant \varepsilon ||A(y^0 - y)||_{B^{-1}}$$

достаточно $k > k_0(\varepsilon) = \dfrac{\ln\frac{2}{\varepsilon}}{2\sqrt{2}\,\sqrt[4]{\eta}}$ итераций.

Замечание. Доказательство теоремы приведено в [2].

Применение к модельной задаче.

Воспользуемся этой теоремой для оценки числа итераций $k_0(\varepsilon)$ в случае применения попеременно-треугольного метода с чебышевскими итерационными параметрами для решения модельной задачи. Для модельной задачи $\delta = \frac{4}{h^2}\sin^2\frac{\pi h}{2}$ и $\Delta = \frac{4}{h^2}\cos^2\frac{\pi h}{2}$.

Тогда, при $\varepsilon = 0,5 \cdot 10^{-4} \approx e^{-10}$ для числа итераций $k_0(\varepsilon)$ справедлива приближенная оценка

$$k_0(\varepsilon) = \frac{\ln 2 + 10}{2\sqrt{2}\sqrt[4]{\mathrm{tg}^2(\frac{\pi h}{2})}} \approx \frac{\ln 2 + 10}{2\sqrt{2}\sqrt{\frac{\pi h}{2}}} = \{\mathsf{h} = \frac{\mathsf{1}}{\mathsf{N}}\} = \frac{(\ln 2 + 10)\sqrt{N}}{2\sqrt{\pi}} \approx 3\sqrt{N}.$$

Таким образом, при решении системы с числом уравнений $N = 10$ понадобится 10 итераций, а в случае $N = 100$ — 30 итераций.

Приведем асимптотические оценки при $N \to \infty$ минимального числа итераций $k_0(\varepsilon)$, необходимого для достижения заданной точности ε, для рассмотренных итерационных методов решения модельной задачи:

$$k_0(\varepsilon) \approx \frac{2N^2}{\pi^2}\ln\frac{1}{\varepsilon} \qquad \text{— для метода простой итерации;}$$
$$k_0(\varepsilon) \approx \frac{N}{2\pi}\ln\frac{1}{\varepsilon} \qquad \text{— для ПТИМ;}$$
$$k_0(\varepsilon) \approx \frac{\sqrt{N}}{2\sqrt{\pi}}\ln\frac{2}{\varepsilon} \qquad \text{— для ПТИМ с чебышевскими параметрами.}$$

Здесь аббревиатура ПТИМ является сокращенным обозначением попеременно–треугольного итерационного метода. Превосходство в скорости сходимости ПТИМ с чебышевскими параметрами над другими рассмотренными итерационными методами очевидно.

1.5 Итерационные методы вариационного типа

1.5.1 Одношаговые итерационные методы вариационного типа

Рассмотрим нестационарный одношаговый итерационный метод решения системы $Ay = f$ вида

$$B\frac{y^{k+1} - y^k}{\tau_{k+1}} + Ay^k = f,\ k = 0, 1, \ldots. \tag{1.16}$$

Здесь невырожденная матрица B не зависит от номера итерации k.

Пусть, как и ранее, $z^k = y^k - y$ есть погрешность на $k-$ой итерации, удовлетворяющая соотношению

$$z^{k+1} = (E - \tau_{k+1} B^{-1} A) z^k.$$

Пусть D — некоторая матрица, удовлетворяющая условиям $D^T = D > 0$. Будем выбирать итерационные параметры τ_{k+1}, минимизирующие энергетическую норму (см. пункт 1.3.2) погрешности $\|z^{k+1}\|_D$. Такой способ построения итерационного процесса называется *локальной минимизацией.*

Обозначим $w^{k+1} = D^{1/2} z^{k+1}$, тогда

$$\|z^{k+1}\|_D = \sqrt{(Dz^{k+1}, z^{k+1})} = \sqrt{(D^{1/2} z^{k+1}, D^{1/2} z^{k+1})} = \|w^{k+1}\|,$$

и минимизация энергетической нормы $\|z^{k+1}\|_D$ эквивалентна минимизации нормы $\|w^{k+1}\|$. Перепишем уравнение для погрешности в следующей форме

$$D^{1/2} z^{k+1} = D^{1/2}(E - \tau_{k+1} B^{-1} A) D^{-1/2} D^{1/2} z^k \;\Leftrightarrow$$
$$\Leftrightarrow\; w^{k+1} = (E - \tau_{k+1} D^{1/2} B^{-1} A D^{-1/2}) w^k \;\Leftrightarrow\; w^{k+1} = (E - \tau_{k+1} C) w^k.$$

Здесь $C = D^{1/2} B^{-1} A D^{-1/2}$ — невырожденная матрица. Считая что $w^k \neq 0$ (иначе на $k-$ой итерации найдено точное решение), получим

$$\|w^{k+1}\|^2 = ((E - \tau_{k+1} C) w^k, (E - \tau_{k+1} C) w^k) =$$
$$= \|w^k\|^2 + \tau_{k+1}^2 (Cw^k, Cw^k) - 2\tau_{k+1} (Cw^k, w^k) =$$
$$= \|w^k\|^2 + (Cw^k, Cw^k) \left(\tau_{k+1}^2 - 2\tau_{k+1} \frac{(Cw^k, w^k)}{(Cw^k, Cw^k)} \right) =$$
$$= \|w^k\|^2 + (Cw^k, Cw^k) \left(\tau_{k+1} - \frac{(Cw^k, w^k)}{(Cw^k, Cw^k)} \right)^2 - \frac{(Cw^k, w^k)^2}{(Cw^k, Cw^k)}.$$

Отсюда следует, что минимальное значение $\|w^{k+1}\|$ достигается при

$$\tau_{k+1} = \frac{(Cw^k, w^k)}{(Cw^k, Cw^k)}, \text{ где } \tau_{k+1} > 0, \quad \text{если } C > 0.$$

В дальнейшем будем предполагать условие $C > 0$ выполненным. Отметим, что при таком выборе τ_{k+1} верно равенство

$$\|w^{k+1}\| = \rho_{k+1} \|w^k\|, \text{ где } \rho_{k+1}^2 = 1 - \frac{(Cw^k, w^k)^2}{(Cw^k, Cw^k)(w^k, w^k)} < 1.$$

Учитывая, что $Cw^k = D^{1/2} B^{-1} A z^k$ и $w^k = D^{1/2} z^k$, получим

$$\tau_{k+1} = \frac{(DB^{-1} A z^k, z^k)}{(DB^{-1} A z^k, B^{-1} A z^k)}.$$

Перепишем уравнение итерационного метода (1.16) в виде

$$y^{k+1} = y^k - \tau_{k+1} B^{-1}(Ay^k - f) = y^k - \tau_{k+1} v^k.$$

Здесь вектор $v^k = B^{-1}(Ay^k - f) = B^{-1}r^k$ называется *поправкой*, а вектор $r^k = Ay^k - f = Az^k$ — *невязкой* на $k-$ой итерации. Поскольку $v^k = B^{-1}Az^k$, приходим к равенству

$$\tau_{k+1} = \frac{(Dv^k, z^k)}{(Dv^k, v^k)}. \tag{1.17}$$

Это выражение содержит погрешность z^k, которую нельзя вычислить, поскольку неизвестно точное решение y задачи $Ay = f$. Однако за счет выбора матрицы D можно выразить τ_{k+1} через значения v^k и r^k, которые могут быть вычислены на каждой итерации. Например, если $A^T = A > 0$, то можно выбрать матрицу $D = A$. В этом случае

$$\tau_{k+1} = \frac{(Av^k, z^k)}{(Av^k, v^k)} = \frac{(v^k, Az^k)}{(Av^k, v^k)} = \frac{(v^k, r^k)}{(Av^k, v^k)}.$$

Таким образом, путем выбора матриц B и D можно строить различные одношаговые итерационные методы вариационного типа. Далее будем называть такие методы *градиентными*.

Замечание. Рассмотрим выражение для итерационного параметра τ_1. Согласно (1.17) в градиентных методах

$$\tau_1 = \frac{(Dv^0, z^0)}{(Dv^0, v^0)} = \frac{(DB^{-1}r^0, z^0)}{(DB^{-1}r^0, B^{-1}r^0)} = \frac{(DB^{-1}Az^0, z^0)}{(DB^{-1}Az^0, B^{-1}Az^0)}.$$

Отметим, что если в градиентных методах в качестве матрицы B взять матрицу A, определяющую решаемую систему линейных алгебраических уравнений, то получим, что

$$\tau_1 = \frac{(Dz^0, z^0)}{(Dz^0, z^0)} = 1.$$

Тогда для итерационного приближения y^1 верно уравнение

$$A\frac{y^1 - y^0}{1} + Ay^0 = f \;\Leftrightarrow\; Ay^1 = f.$$

То есть, y^1 будет совпадать с точным решением исходной системы линейных уравнений. Это означает сходимость градиентных методов с матрицей $B = A$ за одну итерацию к точному решению задачи.

Примеры градиентных методов будут приведены в следующих пунктах. Здесь же будем предполагать, что выбор матриц D и B позволяет вычислять параметры τ_{k+1} на каждой итерации.

Далее обсудим сходимость градиентных методов. Ограничимся так называемым самосопряженным случаем, когда матрица $C^T = C > 0$. Тогда существуют такие положительные постоянные γ_1 и γ_2, что

$$\gamma_1 E \leqslant C \leqslant \gamma_2 E, \text{ где } C = D^{1/2}B^{-1}AD^{-1/2} = D^{-1/2}(DB^{-1}A)D^{-1/2}.$$

Отсюда получим, что требование $C^T = C > 0$ равносильно условиям

$$(DB^{-1}A)^T = DB^{-1}A > 0,\ \gamma_1 D \leqslant DB^{-1}A \leqslant \gamma_2 D,\ \gamma_1 > 0. \tag{1.18}$$

Будем считать, что γ_1 и γ_2 — минимальное и максимальное собственное значение матрицы C, соответственно. Это наиболее точные значения постоянных γ_1 и γ_2, при которых выполнены неравенства в (1.18).

Теорема 1.9. *Пусть выполнены условия* (1.18). *Тогда итерационный метод* (1.16), (1.17) *сходится и для его погрешности справедлива оценка*

$$\|z^k\|_D \leqslant \rho^k \|z^0\|_D,\ \textit{где } \rho = \frac{1-\xi}{1+\xi},\ \xi = \frac{\gamma_1}{\gamma_2}.$$

▼ Доказательство. Напомним, что векторы $w^k = D^{1/2}z^k$ при каждом $k = 0, 1, \ldots$ удовлетворяют соотношению $w^{k+1} = (E - \tau_{k+1}C)w^k$. Сопоставим этому равенству уравнение

$$\frac{w^{k+1} - w^k}{\tau_0} + Cw^k = 0, \text{ где } \tau_0 = \frac{2}{\gamma_1 + \gamma_2},$$

в котором параметр τ_{k+1} заменен значением τ_0. Последнее равенство совпадает с уравнением для погрешности метода простой итерации, примененного для решения линейной системы с матрицей C, где $C^T = C > 0$, и оптимальным параметром τ. Поэтому в силу следствия из теоремы 1.5 справедлива оценка

$$\|w^{k+1}\|\big|_{\tau=\tau_0} \leqslant \rho \|w^k\|.$$

Поскольку в итерационном методе (1.16) параметр τ_{k+1} выбирается исходя из минимизации $\|w^{k+1}\|$, также верно неравенство

$$\|w^{k+1}\|\big|_{\tau=\tau_{k+1}} = \min_{\tau>0} \|w^{k+1}\| \leqslant \|w^{k+1}\|\big|_{\tau=\tau_0}.$$

Приходим к тому, что при каждом $k = 0, 1, \ldots$ справедлива оценка

$$\|w^{k+1}\|\big|_{\tau=\tau_{k+1}} \leqslant \rho \|w^k\| \ \Rightarrow \|z^k\|_D \leqslant \rho^k \|z^0\|_D.$$

▲ Утверждение доказано.

Таким образом показано, что в самосопряженном случае любой градиентный метод сходится не хуже, чем соответствующий ему метод простой итерации с симметричной матрицей C. Это не означает, что скорость градиентного метода не может оказаться выше. Например, если в качестве начального приближения выбрать вектор y^0 такой, что соответствующее ему значение $w^0 = D^{1/2}z^0 = D^{1/2}(y^0 - y) = \mu$, где μ — произвольный собственный вектор матрицы C, отвечающий собственному значению λ, то получим

$$\tau_1 = \frac{(C\mu, \mu)}{(C\mu, C\mu)} = \frac{1}{\lambda},\ \rho_1^2 = 1 - \frac{(C\mu, \mu)^2}{(C\mu, C\mu)(\mu, \mu)} = 0.$$

Отсюда $\|w^1\| = \rho_1\|w^0\| = 0$, что равносильно равенству $y^1 = y$, которое означает сходимость метода за одну итерацию. Таким образом, при «удачном» выборе начального приближения градиентные методы могут иметь существенно более высокую скорость сходимости по сравнению с соответствующими им методами простой итерации.

Замечание. Аналогично можно показать (см. [8]), что при «неудачном» выборе w^0 для всех $k = 0, 1, \ldots$ будет верным равенство $\|w^{k+1}\| = \rho\|w^k\|$. Это означает сходимость со скоростью $\ln(1/\rho)$ и неулучшаемость оценки, доказанной в предыдущей теореме.

Рассмотрим случай, когда $A^T = A > 0$, $B^T = B > 0$, $D = A$. Тогда матрица $DB^{-1}A$ симметрична и неравенства (1.18) принимают вид

$$\gamma_1 A \leqslant AB^{-1}A \leqslant \gamma_2 A \ \Leftrightarrow\ \gamma_1 E \leqslant A^{1/2}B^{-1}A^{1/2} \leqslant \gamma_2 E \ \Leftrightarrow$$
$$\Leftrightarrow\ \gamma_1 A^{-1} \leqslant B^{-1} \leqslant \gamma_2 A^{-1} \ \Leftrightarrow\ \gamma_1 B \leqslant A \leqslant \gamma_2 B.$$

Отсюда вытекает (см. теорему 1.5), что градиентный метод будет сходиться не хуже стационарного метода (1.11) с теми же матрицами A, B и оптимальным итерационным параметром. При этом для задания оптимального параметра необходимо находить минимальное и максимальное собственные значения γ_1 и γ_2 матрицы $B^{-1}A$. Для соответствующего градиентного метода эти данные не требуются, хотя и могут быть использованы для априорных оценок скорости сходимости.

Замечание. Например, если матрица B выбирается так же, как в попеременно–треугольном итерационном методе, для сходимости соответствующего градиентного метода можно получить априорную оценку сходимости на основе теоремы 1.6.

В заключение отметим, что минимальными требованиями, необходимыми для построения градиентных методов, являются условия $D^T = D > 0$, $C = D^{1/2}B^{-1}AD^{-1/2} > 0$, а также возможность нахождения параметров τ_{k+1}

по формуле (1.17). Эти условия не предполагают, что матрица A исходной системы $Ay = f$ непременно должна быть симметричной и(или) положительно определенной.

1.5.2 Примеры одношаговых итерационных методов вариационного типа

Метод скорейшего спуска.

Метод скорейшего спуска применим для случая симметричной и положительно определенной матрицы системы $Ay = f$, то есть $A^T = A > 0$. Тогда можно выбрать матрицу $D = A$ и, как уже отмечалось,

$$\tau_{k+1} = \frac{(Av^k, z^k)}{(Av^k, v^k)} = \frac{(v^k, Az^k)}{(Av^k, v^k)} = \frac{(v^k, r^k)}{(Av^k, v^k)}.$$

Условие $C > 0$ приводит к ограничению $B > 0$, поскольку

$$(Cy, y) = (A^{1/2}B^{-1}A^{1/2}y, y) = (B^{-1}A^{1/2}y, A^{1/2}y) > 0 \ \Rightarrow B > 0.$$

Пусть $B = E$, тогда поправка $v^k = B^{-1}r^k$ совпадает с невязкой $r^k = Ay^k - f$, метод является явным и расчетные формулы принимают вид

$$y^{k+1} = y^k - \tau_{k+1}r^k, \ \tau_{k+1} = \frac{(r^k, r^k)}{(Ar^k, r^k)}, \ k = 0, 1, \ldots.$$

Метод минимальных невязок.

В *методе минимальных невязок* выбирается $D = A^TA$, тем самым $D^T = D > 0$. Итерационные параметры вычисляются следующим образом:

$$\tau_{k+1} = \frac{(A^TAv^k, z^k)}{(A^TAv^k, v^k)} = \frac{(Av^k, Az^k)}{(Av^k, Av^k)} = \frac{(Av^k, r^k)}{(Av^k, Av^k)}.$$

Условие $C > 0$ приводит к следующему ограничению применимости метода:

$$\begin{aligned}&(Cy, y) = (D^{1/2}B^{-1}AD^{-1/2}y, y) = \{\bar{y} = D^{-1/2}y\} = (D^{1/2}B^{-1}A\bar{y}, D^{1/2}\bar{y}) = \\ &= (DB^{-1}A\bar{y}, \bar{y}) = (A^TAB^{-1}A\bar{y}, \bar{y}) = (AB^{-1}A\bar{y}, A\bar{y}) = \{\tilde{y} = B^{-1}A\bar{y}\} = \\ &= (A\tilde{y}, B\tilde{y}) = (B^TA\tilde{y}, \tilde{y}) > 0 \ \Rightarrow B^TA > 0.\end{aligned}$$

Если матрица $A > 0$, можно использовать явный метод, выбирая $B = E$. В противном случае необходимо подбирать легко обратимую матрицу B, для которой условие $B^TA > 0$ выполняется.

Название данного метода объясняется тем, что энергетическая норма погрешности

$$\|z^{k+1}\|_D = \sqrt{(A^T A z^{k+1}, z^{k+1})} = \sqrt{(Az^{k+1}, Az^{k+1})} = \|r^{k+1}\|.$$

Поэтому минимизация нормы $\|z^{k+1}\|_D$ в рассматриваемом методе эквивалентна минимизации нормы невязки $\|r^{k+1}\|$.

Метод минимальных поправок.

В *методе минимальных поправок* выбирается $D = A^T B^{-1} A$. Условие $D^T = D > 0$ приводит к ограничениям на выбор матрицы B, а именно $B^T = B > 0$. Итерационные параметры вычисляются следующим образом:

$$\tau_{k+1} = \frac{(A^T B^{-1} A v^k, z^k)}{(A^T B^{-1} A v^k, v^k)} = \frac{(Av^k, B^{-1} A z^k)}{(B^{-1} A v^k, A v^k)} = \frac{(Av^k, v^k)}{(B^{-1} A v^k, A v^k)}.$$

Условие $C > 0$ приводит к дополнительному ограничению:

$$(Cy, y) = (D^{1/2} B^{-1} A D^{-1/2} y, y) = \{\bar{y} = D^{-1/2} y\} = (D^{1/2} B^{-1} A \bar{y}, D^{1/2} \bar{y}) =$$
$$= (DB^{-1} A \bar{y}, \bar{y}) = (A^T B^{-1} A B^{-1} A \bar{y}, \bar{y}) = (AB^{-1} A \bar{y}, B^{-1} A \bar{y}) =$$
$$= \{\tilde{y} = B^{-1} A \bar{y}\} = (A\tilde{y}, \tilde{y}) > 0 \ \Rightarrow A > 0.$$

Итак, метод применим для случая положительно определенной матрицы A.

Для энергетической нормы погрешности справедливы равенства

$$\|z^{k+1}\|_D^2 = (A^T B^{-1} A z^{k+1}, z^{k+1}) = (B^{-1} r^{k+1}, r^{k+1}) = (v^{k+1}, B v^{k+1}) = \|v^{k+1}\|_B^2.$$

Поэтому минимизация нормы $\|z^{k+1}\|_D$ в рассматриваемом методе эквивалентна минимизации нормы поправки $\|v^{k+1}\|_B$. Отсюда и название метода.

Метод минимальных погрешностей.

Метод минимальных погрешностей определяется следующим выбором матриц D и B:

$$D = B_0, \ B = (A^T)^{-1} B_0, \text{ где } B_0^T = B_0 > 0.$$

Здесь в качестве B_0 можно выбрать произвольную симметричную положительно определенную матрицу, которая легко обратима, например, диагональную. Итерационные параметры вычисляются следующим образом

$$\tau_{k+1} = \frac{(B_0 v^k, z^k)}{(B_0 v^k, v^k)} = \frac{(B_0 B^{-1} r^k, z^k)}{(B_0 B^{-1} r^k, v^k)} = \frac{(B_0 B_0^{-1} A^T r^k, z^k)}{(B_0 B_0^{-1} A^T r^k, v^k)} =$$

$$= \frac{(r^k, Az^k)}{(r^k, Av^k)} = \frac{(r^k, r^k)}{(r^k, Av^k)}.$$

Проверим выполнение условия $C > 0$:

$$(Cy, y) = (D^{1/2}B^{-1}AD^{-1/2}y, y) = \{\bar{y} = D^{-1/2}y\} = (D^{1/2}B^{-1}A\bar{y}, D^{1/2}\bar{y}) =$$
$$= (DB^{-1}A\bar{y}, \bar{y}) = (B_0B_0^{-1}A^TA\bar{y}, \bar{y}) = (A\bar{y}, A\bar{y}) > 0.$$

Метод применим для произвольной невырожденной матрицы A.

Название метода объясняется тем, что на каждой итерации минимизируется норма погрешности $\|z^{k+1}\|_D = \|z^{k+1}\|_{B_0}$.

1.6 Методы сопряженных направлений

Методы решения систем линейных алгебраических уравнений, в которых вектор неизвестных представляется в виде линейной комбинации векторов, сопряженных (ортогональных) в какой-либо метрике, связанной с матрицей решаемой системы уравнений, называют методами сопряженных направлений [7]. Одним из таких методов является метод сопряженных градиентов.

1.6.1 Метод сопряженных градиентов

Пусть невырожденная квадратная матрица $A = (a_{ij})$ $(i, j = 1, 2, \ldots, n)$ является симметричной и положительно определенной $\left(A = A^T > 0\right)$. По-прежнему рассматривается решение системы линейных алгебраических уравнений

$$Ay = f, \tag{1.19}$$

где $f = \begin{pmatrix} f_1 & f_2 & \ldots & f_n \end{pmatrix}^T$ заданный вектор $(f \neq 0)$, а вектор

$y = \begin{pmatrix} y_1 & y_2 & \ldots & y_n \end{pmatrix}^T$ - искомое решение системы уравнений (1.19).

Как уже отмечалось, при $A = A^T > 0$ билинейный функционал

$(Au, v) = (u, Av)$ удовлетворяет всем аксиомам скалярного произведения, что позволяет использовать обозначения $(u, v)_A := (Au, v)$ и $\|u\|_A := \sqrt{(u, u)_A}$ для подчиненной этому скалярному произведению нормы.

Пусть некоторые n-мерные векторы $b^0, b^1, \ldots, b^{n-1}$ являются линейно независимыми. Используя эти векторы, по аналогии с процедурой ортогонализации Грамма-Шмидта (см. [5]), построим векторы $e^1, e^2, \ldots, e^n$ следующим образом:

$$e^1 = \frac{s^1}{\|s^1\|_A}, \text{ где } s^1 = b^0;$$

$$e^2 = \frac{s^2}{\|s^2\|_A}, \text{ где } s^2 = b^1 - \left(b^1, e^1\right)_A e^1;$$

$$e^k = \frac{s^k}{\|s^k\|_A}, \text{ где } s^k = b^{k-1} - \sum_{m=1}^{k-1} \left(b^{k-1}, e^m\right)_A e^m;\ k = 3, 4, \ldots, n. \qquad (1.20)$$

Покажем по индукции, что система векторов $e^1, e^2, \ldots, e^n$ является *A-ортонормированной*, то есть $\left(e^i, e^j\right)_A = \delta_{ij}$, где δ_{ij} – символ Кронекера $(i, j = 1, 2, \ldots, n)$, а, следовательно, система векторов $s^1, s^2, \ldots, s^n$ является

A-ортогональной, то есть $\left(s^i, s^j\right)_A = 0$ при $i \neq j$.

По построению $\left\|e^i\right\|_A = 1, (i = 1, 2, \ldots, n)$. Кроме того (базис индукции)

$$\left(e^2, e^1\right)_A = \frac{1}{\|s^2\|_A} \left(b^1 - \left(b^1, e^1\right)_A e^1, e^1\right)_A =$$

$$= \frac{1}{\|s^2\|_A} \left[\left(b^1, e^1\right)_A - \left(b^1, e^1\right)_A \underbrace{\left\|e^1\right\|_A^2}_{=1}\right] = 0.$$

Пусть система векторов $e^1, e^2, \ldots, e^k$ является A-ортонормированной (предположение индукции). Покажем, что тогда вектор e^{k+1} A-ортогонален каждому вектору этой системы (шаг индукции). Для произвольного $i \in \{1, 2, \ldots, k\}$

$$\left(e^{k+1}, e^i\right)_A = \frac{1}{\|s^{k+1}\|_A} \left(b^k - \sum_{m=1}^{k} \left(b^k, e^m\right)_A e^m, e^i\right)_A =$$

$$= \frac{1}{\|s^{k+1}\|_A} \left[\left(b^k, e^i\right)_A - \left(b^k, e^i\right)_A \underbrace{\left\|e^i\right\|_A^2}_{=1} - \sum_{\substack{m=1 \\ m \neq i}}^{k} \left(b^k, e^m\right)_A \underbrace{\left(e^m, e^i\right)_A}_{=0}\right] = 0.$$

Таким образом, доказано, что система векторов $e^1, e^2, \ldots, e^n$ является A-ортонормированной, откуда вытекает линейная независимость этих векторов. Действительно, если их линейная комбинация $\alpha_1 e^1 + \alpha_2 e^2 + \cdots + \alpha_n e^n = 0$, то после скалярного умножения этого равенства на вектор Ae^i получим, что $\alpha_i = 0$ для любого $i = 1, 2, \ldots, n$.

Линейная независимость системы векторов $e^1, e^2, \ldots, e^n$ позволяет искать решение задачи (1.19) в виде $y = \sum_{i=1}^{n} c_i e^i$, где c_i числовые коэффициенты. Подставляя указанное представление в уравнение (1.19) и умножая полученное равенство скалярно на векторы e^j $(j = 1, 2, \ldots, n)$, получим

$$\sum_{i=1}^{n} c_i \left(e^j, Ae^i\right) = \sum_{i=1}^{n} c_i \left(e^j, e^i\right)_A = \sum_{i=1}^{n} c_i \delta_{ij} = c_j = \left(e^j, f\right).$$

То есть, решением уравнения (1.19) является вектор $y = \sum\limits_{i=1}^{n} \left(e^i, f\right) e^i$.

Введём в рассмотрение вспомогательные векторы

$$y^k = \sum_{i=1}^{k} \left(e^i, f\right) e^i,\ k = 1, 2, \ldots, n,$$

которые также можно определить рекуррентным соотношением

$$y^k = y^{k-1} + \left(e^k, f\right) e^k,\ k = 1, 2, \ldots, n,\ y^0 = 0. \tag{1.21}$$

Тогда решение уравнения (1.19) $y = y^n$. Умножая соотношение (1.21) на матрицу A слева и вычитая из обеих частей полученного равенства вектор f, получим

$$Ay^k - f = Ay^{k-1} - f + \left(e^k, f\right) Ae^k.$$

Вектор $r^k = Ay^k - f$ представляет собой невязку приближенного решения y^k системы уравнений (1.19). При этом $r^0 = Ay^0 - f = -f$. Таким образом, для невязок r^k верно рекуррентное соотношение

$$r^k = r^{k-1} - \left(e^k, r^0\right) Ae^k,\ k = 1, 2, \ldots, n-1,\ r^n = 0,\ r^0 = -f. \tag{1.22}$$

Предполагая линейную независимость системы векторов $r^0, r^1, \ldots, r^{n-1}$ (доказана далее), выберем эти векторы в качестве $b^0, b^1, \ldots, b^{n-1}$ в формулах (1.20). Тогда

$$s^k = r^{k-1} - \sum_{m=1}^{k-1} \left(r^{k-1}, e^m\right)_A e^m,\ k = 2, 3, \ldots, n,\ s^1 = r^0,\ e^k = \frac{s^k}{\left\|s^k\right\|_A}. \tag{1.23}$$

Докажем, что векторы r^k и e^k, задаваемые рекуррентными соотношениями (1.22), (1.23), обладают следующими свойствами:

1. Верны равенства

$$\left(e^k, r^j\right) = \left(e^k, r^0\right),\ k = 1, 2, \ldots, n,\ j = 0, 1, \ldots, k-1 \tag{1.24}$$

и условия ортогональности

$$\left(e^k, r^k\right) = 0,\ k = 1, 2, \ldots, n; \tag{1.25}$$

2. Выполнены условия ортогональности

$$\left(r^k, r^j\right) = 0,\ k = 1, 2, \ldots, n-1,\ j = 0, 1, \ldots, k-1 \tag{1.26}$$

и

$$\left(r^k, e^j\right)_A = 0,\ k = 2, 3, \ldots, n-1,\ j = 1, 2, \ldots, k-1. \tag{1.27}$$

Замечание. 1. Выполнение (1.26) означает, что векторы $r^0,\ r^1,\ \ldots, r^{n-1}$ попарно ортогональны и, следовательно, линейно независимы. 2. Из (1.27) следует, что вектор r^k $(k = 1, 2, \ldots, n-1)$ A-ортогонален любой линейной комбинации векторов e^j $(j = 1, 2, \ldots, k-1)$ и соотношение (1.23) для векторов s^k $(k = 1, 2, \ldots, n)$ упрощается и принимает вид

$$s^k = r^{k-1} - \left(r^{k-1}, e^{k-1}\right)_A e^{k-1},\ k = 2, 3, \ldots, n,\ s^1 = r^0,\ e^k = \frac{s^k}{\|s^k\|_A}. \tag{1.28}$$

Установим справедливость первой группы свойств. При $j < k$

$$\left(e^k, r^j\right) = \left(e^k, r^{j-1} - \left(e^j, r^0\right) Ae^j\right) = \left(e^k, r^{j-1}\right) = \cdots = \left(e^k, r^0\right).$$

Учитывая данное равенство,

$$\left(e^k, r^k\right) = \left(e^k, r^{k-1} - \left(e^k, r^0\right) Ae^k\right) = \left(e^k, r^{k-1}\right) - \left(e^k, r^0\right) =$$
$$= \left(e^k, r^0\right) - \left(e^k, r^0\right) = 0.$$

Вторую группу свойств докажем методом математической индукции (используя соотношения (1.22), (1.23) и (1.25)). Проверим их выполнение для $k = 2$ (базис индукции):

$$\left(r^1, r^0\right) = \left(r^0 - \left(e^1, r^0\right) Ae^1, r^0\right) = \left(r^0, r^0\right) - \left(r^0, r^0\right) = 0;$$
$$\left(r^2, r^0\right) = \left(r^1 - \left(e^2, r^0\right) Ae^2, r^0\right) = -\left(e^2, r^0\right)\left(Ae^2, \left\|s^1\right\|_A e^1\right) = 0;$$
$$\left(r^2, r^1\right) = \left(r^2, \left\|s^2\right\|_A e^2 + \left(r^1, e^1\right)_A e^1\right) = \left(r^1, e^1\right)_A \left(r^2, e^1\right) =$$
$$= \left(r^1, e^1\right)_A \left(r^1 - \left(e^2, r^0\right) Ae^2, e^1\right) = 0;$$
$$\left(r^2, e^1\right)_A = \left(r^2, Ae^1\right) = \left(r^2, \frac{r^0 - r^1}{\left(e^1, r^0\right)}\right) = 0.$$

Знаменатель $\left(e^1, r^0\right) = \left(\frac{r^0}{\|r^0\|_A}, r^0\right) = \left\{r^0 = -f,\ f \neq 0\right\} \neq 0.$

Пусть вторая группа свойств справедлива для некоторого $k \geq 2$ (предположение индукции).

Достаточно доказать (шаг индукции), что

$$\left(r^{k+1}, r^j\right) = 0,\ j = 0, 1, \ldots, k;$$
$$\left(r^{k+1}, e^j\right)_A = 0,\ j = 1, 2, \ldots, k.$$

Покажем справедливость этих равенств:

$$\begin{aligned}
&\left(r^{k+1}, r^j\right) = \left(r^k - \left(e^{k+1}, r^0\right) Ae^{k+1}, r^j\right) = -\left(e^{k+1}, r^0\right)\left(Ae^{k+1}, r^j\right) = \\
&= -\left(e^{k+1}, r^0\right)\left(Ae^{k+1}, \left\|s^{j+1}\right\|_A e^{j+1} + \left(r^j, e^j\right)_A e^j\right) = 0, \text{ для } j < k; \\
&\left(r^{k+1}, r^k\right) = \left(r^k - \left(e^{k+1}, r^0\right) Ae^{k+1}, \left\|s^{k+1}\right\|_A e^{k+1} + \left(r^k, e^k\right)_A e^k\right) = \\
&= \left\|s^{k+1}\right\|_A \left(r^k, e^{k+1}\right) - \left\|s^{k+1}\right\|_A \left(e^{k+1}, r^0\right) = 0, \text{ для } j = k; \\
&\left(r^{k+1}, e^j\right)_A = \left(r^{k+1}, Ae^j\right) = \left(r^{k+1}, \frac{r^{j-1} - r^j}{\left(e^j, r^0\right)}\right) = 0.
\end{aligned}$$

Замечание. Если $\left(e^j, r^0\right) = 0$, то $r^j = r^{j-1}$. Тогда, условие ортогональности $\left(r^j, r^{j-1}\right) = 0$ выполнено только при $r^{j-1} = 0$. Это означает, что решение задачи (1.19) $y = y^{j-1}$ найдено на $(j-1)$-ом шаге применения рекуррентной формулы (1.21) (в совокупности с (1.22), (1.23)). В этом случае выполнения свойства (1.27) для $k+1 > j$ не требуется.

Из соотношений (1.21), (1.22), (1.28) следует, что вектор y, являющийся решением системы линейных алгебраических уравнений (1.19), может быть вычислен по следующему алгоритму.

Пусть $y^0 = 0$. Тогда $r^0 = Ay^0 - f = -f$. Пусть $s^1 = r^0$.

Далее, последовательно, вычисляются

$$\begin{aligned}
y^k &= y^{k-1} - \frac{\left(s^k,\ r^0\right)}{\left(s^k,\ As^k\right)} s^k, \\
r^k &= r^{k-1} - \frac{\left(s^k,\ r^0\right)}{\left(s^k,\ As^k\right)} As^k \ \text{ или } \ r^k = Ay^k - f \ \text{ для } \ k = 1, 2, \ldots, n \ , \\
s^{k+1} &= r^k - \frac{\left(r^k,\ As^k\right)}{\left(s^k,\ As^k\right)} s^k \ \text{ для } \ k = 1, 2, \ldots, (n-1).
\end{aligned} \tag{1.29}$$

Вектор $y^n = y$ является искомым решением системы (1.19).

Преобразуем формулы (1.29) к виду, который, как показала практика вычислений, менее чувствителен к ошибкам машинного округления чисел. Из (1.24) и (1.28) следует, что при $k = 1, 2, \ldots, n$

$$\begin{aligned}
\left(s^k,\ r^0\right) = \left(s^k,\ r^{k-1}\right) &= \left(r^{k-1} - \frac{\left(r^{k-1},\ As^{k-1}\right)}{\left(s^{k-1},\ As^{k-1}\right)} s^{k-1},\ r^{k-1}\right) = \\
&= \left(r^{k-1},\ r^{k-1}\right) - \frac{\left(r^{k-1},\ As^{k-1}\right)}{\left(s^{k-1},\ As^{k-1}\right)} \underbrace{\left(s^{k-1},\ r^{k-1}\right)}_{=0 \ (1.25)} = \left(r^{k-1},\ r^{k-1}\right).
\end{aligned} \tag{1.30}$$

Рассмотрим скалярное произведение $\left(r^k,\ r^k\right)$. В силу (1.22), (1.26) и (1.30), имеем

$$\left(r^k,\ r^k\right) = \left(r^{k-1} - \frac{\left(s^k,\ r^0\right)}{\left(s^k,\ As^k\right)} As^k,\ r^k\right) = \underbrace{\left(r^{k-1},\ r^k\right)}_{=0\ (1.26)} -$$

$$- \frac{\left(s^k,\ r^0\right)}{\left(s^k,\ As^k\right)} \left(As^k,\ r^k\right) = -\frac{\left(r^{k-1},\ r^{k-1}\right)}{\left(s^k,\ As^k\right)} \left(As^k,\ r^k\right),$$

где $k = 1, 2, \ldots, n$.

Следовательно

$$\frac{\left(r^k,\ As^k\right)}{\left(s^k,\ As^k\right)} = -\frac{\left(r^k,\ r^k\right)}{\left(r^{k-1},\ r^{k-1}\right)} \tag{1.31}$$

при $k = 1, 2, \ldots, n$.

Введём в рассмотрение вектор p^k, связанный с вектором s^{k+1} соотношением (см. [11])

$$s^{k+1} = p^k \left(r^k,\ r^k\right) \tag{1.32}$$

при $k = 0, 1, \ldots, (n-1)$.

Вектора $p^0,\ p^1,\ \ldots, p^{n-1}$ являются A-ортогональными векторами.

В результате подстановки (1.30), (1.31) и (1.32) в (1.29) получим следующие рассчётные формулы метода сопряжённых градиентов:

$$\begin{aligned}
&y^0 = 0,\ \ r^0 = -f,\ \ p^0 = \frac{r^0}{\left(r^0,\ r^0\right)},\ \ \text{далее} \\
&y^k = y^{k-1} - \frac{p^{k-1}}{\left(p^{k-1},\ Ap^{k-1}\right)}, \\
&r^k = r^{k-1} - \frac{Ap^{k-1}}{\left(p^{k-1},\ Ap^{k-1}\right)} \ \text{ или } \ r^k = Ay^k - f \\
&\qquad \text{для } \ k = 1, 2, \ldots, n\ , \\
&p^k = p^{k-1} + \frac{r^k}{\left(r^k,\ r^k\right)} \ \text{ для } \ k = 1, 2, \ldots, (n-1).
\end{aligned} \tag{1.33}$$

При реализации алгоритма (1.33) требуется выполнить $O(n^3)$ операций умножения.

Минимизационные свойства метода сопряженных градиентов

Пусть вектор $y = (y_1,\ \ldots\ ,\ y_n)^T$ является искомым решением системы линейных алгебраических уравнений $Ay = f$ (1.19), где матрица $A = A^T > 0$, а $x = (x_1,\ \ldots\ ,\ x_n)^T$ - произвольный вектор. Вектор $z = x - y$ является ошибкой (погрешностью) определения вектора y. При наличии положительной

определенности у матрицы A, удобной количественной мерой точности решения системы (1.19) является функция многих переменных $f(z) = (Az,\ z) \geq 0$, которую называют функцией ошибок.

Функция ошибок принимает положительные значения при любых $x \neq y$ и имеет минимальное значение равное нулю при $x = y$. Следовательно, решение системы (1.19) эквивалентно поиску вектора x, при котором функция ошибок принимает минимальное значение равное нулю.

Преобразуем выражение для функции ошибок к виду

$$\begin{aligned} f(z) &= (Az,\ z) = (Ax - Ay,\ x - y) = \{Ay = f\} = \\ &= (Ax - f,\ x - y) = (Ax,\ x) - (Ax,\ y) - (f,\ x) + (f,\ y) = \\ &= \{(Ax,\ y) = (x,\ Ay) = (x,\ f)\} = (Ax,\ x) - 2(f,\ x) + (f,\ y) = \\ &= J(x) + (f,\ y), \end{aligned}$$

где функция $J(x) = (Ax,\ x) - 2(f,\ x)$.

Итак, решение системы линейных алгебраических уравнений (1.19) эквивалентно поиску вектора x, доставляющего минимум функции
$J(x) = (Ax,\ x) - 2(f,\ x) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j - 2 \sum_{i=1}^{n} f_i x_i$.

Рассмотрим вектор $\operatorname{grad} J(x) = \left(\frac{\partial J}{\partial x_1},\ \frac{\partial J}{\partial x_2},\ \ldots,\ \frac{\partial J}{\partial x_n}\right)^T$, определяющий направление самого быстрого роста значений функции $J(x)$. Компонента $\frac{\partial J(x)}{\partial x_l}$ $(l = 1, 2, \ldots, n)$ этого вектора равна

$$\begin{aligned} &\frac{\partial J(x_1,\ \ldots\ ,\ x_n)}{\partial x_l} = \frac{\partial}{\partial x_l} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j - 2 \sum_{i=1}^{n} f_i x_i \right) = \\ &= \frac{\partial}{\partial x_l} \sum_{i=1}^{n} x_i \left(\sum_{j=1}^{l-1} a_{ij} x_j + a_{il} x_l + \sum_{j=l+1}^{n} a_{ij} x_j - 2 f_i \right) = \\ &= \frac{\partial}{\partial x_l} \left(\sum_{i=1}^{l-1} x_i \left(\sum_{j=1}^{l-1} a_{ij} x_j + a_{il} x_l + \sum_{j=l+1}^{n} a_{ij} x_j - 2 f_i \right) + \right. \\ &+ x_l \left(\sum_{j=1}^{l-1} a_{lj} x_j + a_{ll} x_l + \sum_{j=l+1}^{n} a_{lj} x_j - 2 f_l \right) + \sum_{i=l+1}^{n} x_i \left(\sum_{j=1}^{l-1} a_{ij} x_j + \right. \\ &\left. \left. + a_{il} x_l + \sum_{j=l+1}^{n} a_{ij} x_j - 2 f_i \right) \right) = \frac{\partial}{\partial x_l} \left(x_l \sum_{i=1}^{l-1} a_{il} x_i + x_l \sum_{j=1}^{l-1} a_{lj} x_j + \right. \\ &\left. + x_l a_{ll} x_l + x_l \sum_{j=l+1}^{n} a_{lj} x_j - x_l 2 f_l + x_l \sum_{i=l+1}^{n} x_i a_{il} \right) = \sum_{i=1}^{l-1} a_{il} x_i + \end{aligned}$$

$$+\sum_{j=1}^{l-1} a_{lj}x_j + 2a_{ll}x_l + \sum_{j=l+1}^{n} a_{lj}x_j - 2f_l + \sum_{i=l+1}^{n} a_{il}x_i = \sum_{i=1}^{n} a_{il}x_i +$$
$$+\sum_{j=1}^{n} a_{lj}x_j - 2f_l = \left\{A = A^T\right\} = 2\left(\sum_{j=1}^{n} a_{lj}x_j - f_l\right) =$$
$$= 2\left((Ax)_l - f_l\right) = 2\left(Ax - f\right)_l = 2r_l,$$

где $r = Ax - f$ вектор невязки в системе линейных алгебраических уравнений (1.19).

То есть, вектор $\operatorname{grad} J(x) = 2r$. Следовательно, вектор $(-r)$ определяет направление самого быстрого убывания значений функции J в точке x.

Рассмотрим функцию J от аргумента $x - \tau r$, принадлежащего направлению антиградиента функции $J(x)$ при значениях числового параметра $\tau > 0$:

$$J\left(x - \tau r\right) = \left(A(x - \tau r),\ x - \tau r\right) - 2\left(f,\ x - \tau r\right) = \left(Ax,\ x\right) -$$
$$- 2\tau\left(Ax,\ r\right) + \tau^2\left(Ar,\ r\right) - 2\left(f,\ x\right) + 2\tau\left(f,\ r\right) = \left(Ax,\ x\right) - 2\left(f,\ x\right) -$$
$$- 2\tau\left(r,\ r\right) + \tau^2\left(Ar,\ r\right) = J\left(x\right) + \left(Ar,\ r\right)\left(\tau^2 - 2\tau\frac{\left(r,\ r\right)}{\left(Ar,\ r\right)}\right) =$$
$$= J\left(x\right) + \left(Ar,\ r\right)\left(\tau - \frac{\left(r,\ r\right)}{\left(Ar,\ r\right)}\right)^2 - \frac{\left(r,\ r\right)^2}{\left(Ar,\ r\right)}.$$

Пусть для параметра τ выполнено условие

$$\tau = \frac{\left(r,\ r\right)}{\left(Ar,\ r\right)}. \tag{1.34}$$

Тогда функция J в точке $x - \frac{(r,\,r)}{(Ar,\,r)}r$ будет иметь минимальное значение вдоль направления $(-r)$.

Выражение (1.34) совпадает с формулой для итерационного параметра τ_{k+1} в методе скорейшего спуска (см. п.1.5.2).

Рассмотрим частный случай $x = \left(x_1,\ x_2\right)^T$ $(n = 2)$. При $n = 2$ поверхность $J(x)$ будет иметь вид эллипсоида. Линия $J(x) = const$ представляет собой эллипс, который может быть сильно вытянут вдоль обной из полуосей. Если точка x, являющаяся аргументом функции $J(x)$, расположена близко к концу большей полуоси эллипса, то и следующее приближение, точка $x - \frac{(r,\,r)}{(Ar,\,r)}r$, будет расположено близко к концу большей полуоси эллипса. То есть, выбор направления антиградиента функции $J(x)$ не является оптимальным для приближения к искомому решению y.

В методе сопряжённых градиентов очередное приближение

$y^k = y^{k-1} - \frac{p^{k-1}}{(p^{k-1},\ Ap^{k-1})}$ (1.33) к искомому решению расположено на направлении $\left(-p^{k-1}\right)$, не совпадающем с направлением антиградиента функции J.

Для значения функции $J(y^k)$ верна следующая оценка:

$$
\begin{aligned}
&J\left(y^k\right) = \left(Ay^k,\ y^k\right) - 2\left(f,\ y^k\right) = \left\{y^k = y^{k-1} - \frac{p^{k-1}}{(p^{k-1},\ Ap^{k-1})}\ (1.33)\right\} = \\
&= \left(A\left(y^{k-1} - \frac{p^{k-1}}{(p^{k-1},\ Ap^{k-1})}\right),\ y^{k-1} - \frac{p^{k-1}}{(p^{k-1},\ Ap^{k-1})}\right) - \\
&- 2\left(f,\ y^{k-1} - \frac{p^{k-1}}{(p^{k-1},\ Ap^{k-1})}\right) = \left(Ay^{k-1},\ y^{k-1}\right) - \\
&- \frac{1}{(p^{k-1},\ Ap^{k-1})}\left(Ay^{k-1},\ p^{k-1}\right) - \frac{1}{(p^{k-1},\ Ap^{k-1})}\left(Ap^{k-1},\ y^{k-1}\right) + \\
&+ \frac{1}{(p^{k-1},\ Ap^{k-1})^2}\left(Ap^{k-1},\ p^{k-1}\right) - 2\left(f,\ y^{k-1}\right) + \frac{2}{(p^{k-1},\ Ap^{k-1})}\left(f,\ p^{k-1}\right) = \\
&= J\left(y^{k-1}\right) - \frac{2}{(p^{k-1},\ Ap^{k-1})}\left(Ay^{k-1} - f,\ p^{k-1}\right) + \frac{1}{(p^{k-1},\ Ap^{k-1})} = \\
&= J\left(y^{k-1}\right) - \frac{2}{(p^{k-1},\ Ap^{k-1})}\left(r^{k-1},\ p^{k-1}\right) + \frac{1}{(p^{k-1},\ Ap^{k-1})} = \\
&= J\left(y^{k-1}\right) - \frac{1}{(p^{k-1},\ Ap^{k-1})}\left(2\left(r^{k-1},\ p^{k-1}\right) - 1\right) = \\
&= \left\{\left(r^{k-1},\ p^{k-1}\right) = \left(p^{k-1},\ r^0\right)\ \ (1.24)\right\} = J\left(y^{k-1}\right) - \\
&- \frac{1}{(p^{k-1},\ Ap^{k-1})}\left(2\left(p^{k-1},\ r^0\right) - 1\right) = \left\{2\left(p^{k-1},\ r^0\right) - 1 = \right. \\
&= 2\left(\underbrace{p^{k-2} + \frac{r^{k-1}}{\left(r^{k-1},\ r^{k-1}\right)}}_{(1.33)},\ r^0\right) - 1 = 2\left(p^{k-2},\ r^0\right) + \\
&+ \frac{2}{\left(r^{k-1},\ r^{k-1}\right)}\underbrace{\left(r^{k-1},\ r^0\right)}_{=0\ (1.26)} - 1 = 2\left(p^{k-2},\ r^0\right) - 1 = \ldots = 2\left(p^0,\ r^0\right) - 1 = \\
&\left. = 2\left(\frac{r^0}{(r^0,\ r^0)},\ r^0\right) - 1 = 1\right\} = J\left(y^{k-1}\right) - \frac{1}{(p^{k-1},\ Ap^{k-1})} < J\left(y^{k-1}\right).
\end{aligned}
$$

То есть, алгоритм метода сопряжённых градиентов гарантирует монотонное убывание значений функции J.

1.6.2 Метод Крейга

Метод Крейга применим для решения систем линейных алгебраических уравнений

$$Ay = f \tag{1.35}$$

с невырожденной матрицей A.

Будем искать вектор y, используя вспомогательный вектор x такой, что $y = A^T x$. Тогда исходная система (1.35) примет вид $AA^T x = f$. Матрица AA^T преобразованной системы является симметричной и положительно определенной. Поэтому, для решения преобразованной системы можно использовать метод сопряженных градиентов.

Алгоритм метода сопряжённых градиентов (1.33) для преобразованной системы имеет вид:

$$\begin{aligned}
&x^0 = 0, \;\; r^0 = -f, \;\; p^0 = \frac{r^0}{(r^0,\; r^0)}, \;\; \text{далее} \\
&x^k = x^{k-1} - \frac{p^{k-1}}{(p^{k-1},\; AA^T p^{k-1})}, \\
&r^k = AA^T x^k - f \;\; \text{для} \;\; k = 1, 2, \ldots, n \;\; , \\
&p^k = p^{k-1} + \frac{r^k}{(r^k,\; r^k)} \;\; \text{для} \;\; k = 1, 2, \ldots, (n-1).
\end{aligned}$$

Умножая соотношение для очередного приближения x^k слева на матрицу A^T и учитывая, что $y = A^T x$, получим следующий алгоритм решения системы (1.35) :

$$\begin{aligned}
&y^0 = 0, \;\; r^0 = -f, \;\; p^0 = \frac{r^0}{(r^0,\; r^0)}, \;\; \text{далее} \\
&y^k = y^{k-1} - \frac{A^T p^{k-1}}{(A^T p^{k-1},\; A^T p^{k-1})}, \\
&r^k = Ay^k - f \;\; \text{для} \;\; k = 1, 2, \ldots, n \;\; , \\
&p^k = p^{k-1} + \frac{r^k}{(r^k,\; r^k)} \;\; \text{для} \;\; k = 1, 2, \ldots, (n-1).
\end{aligned} \tag{1.36}$$

Алгоритм (1.36) можно использовать и для решения недоопределенных систем линейных алгебраических уравнений.

1.6.3 Симметризованные сопряжённые градиенты

Метод применим для решения систем линейных алгебраических уравнений $Ay = f$ с невырожденной матрицей A.

Умножим исходную систему $Ay = f$ слева на матрицу A^T. Для решения преобразованной системы $A^T Ay = A^T f$ используем метод сопряжённых градиентов. Вычислительный алгоритм решения преобразованной системы имеет вид:

$$
\begin{aligned}
& y^0 = 0, \quad R^0 = -A^T f, \quad p^0 = \frac{R^0}{(R^0,\ R^0)}, \quad \textit{далее} \\
& y^k = y^{k-1} - \frac{p^{k-1}}{(Ap^{k-1},\ Ap^{k-1})}, \\
& R^k = A^T \left(Ay^k - f \right) \quad \textit{для} \quad k = 1, 2, \ldots, n \quad \textit{и} \\
& p^k = p^{k-1} + \frac{R^k}{(R^k,\ R^k)} \quad \textit{для} \quad k = 1, 2, \ldots, (n-1).
\end{aligned}
\tag{1.37}
$$

Алгоритм (1.37) можно использовать и для решения переопределенных систем линейных алгебраических уравнений.

Глава 2

Задачи на собственные значения

Пусть $A = (a_{ij})$ — матрица размера $n \times n$, а $y = (y_1, y_2, \ldots, y_n)^T$ — вектор неизвестных. Тогда поиск таких констант λ и векторов $y \neq 0$, что

$$Ay = \lambda y,$$

называется задачей на собственные значения. Эта задача эквивалентна поиску таких y, для которых

$$(A - \lambda E)y = 0.$$

При этом λ называют собственными значениями матрицы A, а соответствующие им вектора y — собственными векторами.

Известно, что если $\det(A - \lambda E) = 0$, то решение задачи существует. Этот определитель является полиномом степени n от λ с коэффициентами составленными из элементов матрицы A. Корни полинома легко найти при $n \leq 3$ или если матрица A является диагональной либо треугольной.

Задачу нахождения всех собственных значений матрицы A называют полной проблемой собственных значений. Если нужно найти лишь некоторые λ, то такую задачу называют частичной проблемой собственных значений.

2.1 Поиск собственных значений методом вращений

Метод вращений, предложенный К. Якоби в 1846 году, позволяет найти все собственные значения (решает полную проблему собственных значений) вещественной симметричной матрицы $A = A^T$.

Для матрицы $A = A^T$ справедливо представление вида

$$A = Q^T \Lambda\, Q, \tag{2.1}$$

где Q — ортогональная матрица ($Q^T = Q^{-1}$), а Λ — диагональная матрица, элементами которой являются собственные значения матрицы A, $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$.

Если умножить равенство (2.1) на $(Q^T)^{-1} = Q$ слева и на матрицу $Q^{-1} = Q^T$ справа, то получим

$$QAQ^T = \Lambda.$$

Итак, для нахождения собственных значений матрицы A необходимо построить ортогональную матрицу Q и провести два матричных умножения.

Заметим, что произведение нескольких ортогональных матриц является ортогональной матрицей. Матрицу Q будем строить итерационно, проводя с помощью специальных ортогональных матриц преобразования матрицы A с целью уменьшения абсолютных значений ее недиагональных элементов.

Рассмотрим последовательность матриц

$$\begin{aligned} A_1 &= V_{ij}^1\, A\, (V_{ij}^1)^T, \\ A_2 &= V_{ij}^2\, A_1\, (V_{ij}^2)^T, \ \cdots\ , \\ A_k &= V_{ij}^k\, A_{k-1}\, (V_{ij}^k)^T, \cdots\ . \end{aligned}$$

Здесь V_{ij}^k — ортогональные матрицы следующего вида

$$V_{ij}^k = \begin{pmatrix} 1 & & & & 0 & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ & & & \cos\varphi & & -\sin\varphi & & \\ & & & & 1 & & 0 & \\ 0 & & & & & \ddots & & 0 \\ & & & & 0 & & 1 & \\ & & & \sin\varphi & & \cos\varphi & & \\ & & & & & & & 1 \\ & & & & & & \ddots & \\ & & & & 0 & & & 1 \end{pmatrix}.$$

Элементы матрицы $V_{ij}^k = (v_{lm})$ задаются следующим образом. Диагональные элементы $v_{ll} = 1$ при $l \neq i\,, j$ и $v_{ll} = \cos\varphi$ при $l = i, j$. Вне диагонали $v_{ij} = -\sin\varphi\,$, $v_{ji} = \sin\varphi$, а все остальные элементы равны нулю. Здесь φ — пока свободный параметр. Матрицы V_{ij}^k являются ортогональными матрицами, так как $V_{ij}^k\,(V_{ij}^k)^T = E$. Индекс k — номер итерации. Индексы i и j выбираются на каждой итерации k равными соответствующим индексам максимального по модулю элемента матрицы $A_{k-1} = (a_{lm}^{k-1})$, являющейся $(k-1)$-ым итерационным приближением к матрице Λ. Итак

$$|a_{ij}^{k-1}| = \max_{\substack{l,m \\ l \neq m}} |a_{lm}^{k-1}|\,.$$

Если максимальных по модулю элементов матрицы A_{k-1} несколько, то используется любой из них. Если все недиагональные элементы матрицы A_{k-1} равны нулю, то итерационный процесс построения матриц A_k прекращается.

В качестве количественной характеристики диагональности матрицы A_k выберем число

$$t(A_k) = \sum_{m=1}^{n} \sum_{\substack{s=1 \\ s \neq m}}^{n} (a_{ms}^k)^2 .$$

Если числовая последовательность $t(A_k) \overset{k\to\infty}{\longrightarrow} 0$, то последовательность матриц A_k сходится к диагональной матрице.

Установим соотношения, связывающие элементы матриц A_{k+1} и A_k. Итак

$$A_{k+1} = V_{ij}^{k+1} \, A_k \, (V_{ij}^{k+1})^T.$$

Введем вспомогательные обозначения $B = A_k \, (V_{ij}^{k+1})^T = (b_{ms})$ и $(V_{ij}^{k+1})^T = (\overline{v}_{lm})$. Тогда, по определению произведения матриц,

$$b_{ms} = \sum_{p=1}^{n} a_{mp}^k \, \overline{v}_{ps} = \begin{cases} a_{ms}^k, & s \neq i, j; \\ a_{mi}^k \cos\varphi - a_{mj}^k \sin\varphi, & s = i; \\ a_{mi}^k \sin\varphi + a_{mj}^k \cos\varphi, & s = j. \end{cases} \tag{2.2}$$

То есть, в матрицах B и A_k не совпадают элементы только в столбце с номером i и в столбце с номером j.

Для элементов матрицы $A_{k+1} = V_{ij}^{k+1} B$ верны соотношения

$$a_{ms}^{k+1} = \sum_{p=1}^{n} v_{mp} \, b_{ps} = \begin{cases} b_{ms}, & m \neq i, j; \\ b_{is} \cos\varphi - b_{js} \sin\varphi, & m = i; \\ b_{is} \sin\varphi + b_{js} \cos\varphi, & m = j. \end{cases} \tag{2.3}$$

В матрицах A_{k+1} и B не совпадают элементы только в строке с номером i и в строке с номером j.

Используя (2.3), получаем, что

$$a_{ij}^{k+1} = b_{ij} \cos\varphi - b_{jj} \sin\varphi.$$

Подставляя в это соотношение b_{ij} и b_{jj}, взятые из (2.2), и проводя ряд преобразований приходим к следующему выражению

$$\begin{aligned} a_{ij}^{k+1} &= (a_{ii}^k \sin\varphi + a_{ij}^k \cos\varphi)\cos\varphi - (a_{ji}^k \sin\varphi + a_{jj}^k \cos\varphi)\sin\varphi = \\ &= \{\mathsf{A} = \mathsf{A}^\mathsf{T} \Rightarrow \mathsf{A_k} = \mathsf{A_k^T}\} = (a_{ii}^k - a_{jj}^k)\sin\varphi\cos\varphi + a_{ij}^k(\cos^2\varphi - \sin^2\varphi) = \\ &= \frac{(a_{ii}^k - a_{jj}^k)\sin 2\varphi}{2} + a_{ij}^k \cos 2\varphi. \end{aligned}$$

Элемент a_{ij}^k является максимальным по модулю внедиагональным элементом матрицы A_k. Потребуем выполнения равенства $a_{ij}^{k+1} = 0$. Тогда предыдущее выражение превращается в уравнение относительно φ. Решая его, находим значение параметра φ, используемого для вычисления элементов матрицы V_{ij}^{k+1}

$$\varphi = \frac{1}{2}\operatorname{arctg}\frac{2a_{ij}^k}{a_{jj}^k - a_{ii}^k}.$$

Вычислим количественную характеристику диагональности матрицы A_{k+1}

$$t(A_{k+1}) = \sum_{m=1}^{n}\sum_{\substack{s=1\\ s\neq m}}^{n}(a_{ms}^{k+1})^2.$$

Согласно формулам (2.2) и (2.3), элементы матрицы A_{k+1} отличаются от элементов матрицы A_k только в i-х и j-х строках и столбцах.

Выделим в $t(A_{k+1})$ совпадающие элементы матриц A_{k+1} и A_k в отдельную сумму и проведем следующие преобразования

$$\begin{aligned}
t(A_{k+1}) &= \sum_{\substack{m=1\\ m\neq i,j}}^{n}\sum_{\substack{s=1\\ s\neq i,j,m}}^{n}(a_{ms}^k)^2 + \sum_{\substack{m=1\\ m\neq i,j}}^{n}\left[b_{mi}^2 + b_{mj}^2\right] + \\
&+ \sum_{\substack{s=1\\ s\neq i,j}}^{n}\left[(a_{is}^{k+1})^2 + (a_{js}^{k+1})^2\right] = \sum_{\substack{m=1\\ m\neq i,j}}^{n}\sum_{\substack{s=1\\ s\neq i,j,m}}^{n}(a_{ms}^k)^2 + \\
&+ \sum_{\substack{m=1\\ m\neq i,j}}^{n}\left[(a_{mi}^k)^2\cos^2\varphi + (a_{mj}^k)^2\sin^2\varphi - 2a_{mi}^k a_{mj}^k\sin\varphi\cos\varphi + \right. \\
&\left. + (a_{mi}^k)^2\sin^2\varphi + (a_{mj}^k)^2\cos^2\varphi + 2a_{mi}^k a_{mj}^k\sin\varphi\cos\varphi\right] + \\
&+ \sum_{\substack{s=1\\ s\neq i,j}}^{n}\left[b_{is}^2\cos^2\varphi + b_{js}^2\sin^2\varphi - 2b_{is}b_{js}\sin\varphi\cos\varphi + \right. \\
&\left. + b_{is}^2\sin^2\varphi + b_{js}^2\cos^2\varphi + 2b_{is}b_{js}\sin\varphi\cos\varphi\right] = \\
&= \sum_{\substack{m=1\\ m\neq i,j}}^{n}\sum_{\substack{s=1\\ s\neq i,j,m}}^{n}(a_{ms}^k)^2 + \sum_{\substack{m=1\\ m\neq i,j}}^{n}\left[(a_{mi}^k)^2 + (a_{mj}^k)^2\right] + \sum_{\substack{s=1\\ s\neq i,j}}^{n}\left[(a_{is}^k)^2 + (a_{js}^k)^2\right] + \\
&+ 2(a_{ij}^k)^2 - 2(a_{ij}^k)^2 = t(A_k) - 2(a_{ij}^k)^2.
\end{aligned}$$

То есть $t(A_{k+1}) < t(A_k)$. Уменьшение количественной характеристики диагональности происходит монотонно с ростом номера итерации k на величину равную $2(a_{ij}^k)^2$ — удвоенный квадрат максимального внедиагонального элемента матрицы A_k. Следовательно, последовательность матриц A_k сходится к диагональной матрице.

Получим оценку на количественную характеристику диагональности матрицы A_k. Так как a_{ij}^k — максимальный по модулю внедиагональный элемент, то верно неравенство

$$t(A_k) \leqslant n(n-1)(a_{ij}^k)^2.$$

Отсюда следует, что $(a_{ij}^k)^2 \geqslant \dfrac{t(A_k)}{n(n-1)}$ для $n \geqslant 2$. Подставляя это неравенство в соотношение, связывающее $t(A_k)$ и $t(A_{k+1})$, имеем

$$t(A_{k+1}) = t(A_k) - 2(a_{ij}^k)^2 \leqslant t(A_k) - \frac{2}{n(n-1)} t(A_k) = \rho t(A_k),$$

где $\rho = 1 - \dfrac{2}{n(n-1)} < 1.$

Применив эту оценку k раз, получим

$$t(A_k) \leqslant \rho^k t(A).$$

Итак, последовательность матриц A_k сходится к диагональной матрице Λ со скоростью геометрической прогрессии со знаменателем ρ.

Замечание. Реализация правила выбора индексов i и j в матрице V_{ij}^k требует сравнения $n^2/2$ чисел. Возможна некоторая оптимизация вычислительных затрат. Например, сначала выбирается строка i с максимальной суммой квадратов значений недиагональных элементов. Затем в этой строке выбирается максимальный по модулю элемент a_{ij}^k.

2.2 Степенной метод поиска собственных значений

Рассмотрим задачу поиска максимального по модулю собственного значения симметричной матрицы $A = A^T$ ($\lambda(A)$ — вещественные числа).

Пусть все собственные числа $\lambda(A)$ различны и пронумерованы так, что $|\lambda_1| > |\lambda_2| > |\lambda_3| > \ldots > |\lambda_n|$.

Примечание. Так как нет совпадающих собственных значений, то все собственные вектора $\xi_i, i = 1, \ldots, n$ матрицы A ортогональны и образуют базис, который будем считать ортонормированным - $(\xi_i, \xi_j) = 1$ при $i = j$ и $(\xi_i, \xi_j) = 0$ если $i \neq j$.

Выберем произвольный вектор y^0, отличный от нуля, и построим последовательность векторов y^k

$$y^{k+1} = Ay^k, \quad k = 0, 1, \ldots. \tag{2.4}$$

Используем вектора y^k для вычисления элементов числовой последовательности $\{\Lambda_1^k\}$

$$\Lambda_1^k = \frac{(y^{k+1}, y^k)}{(y^k, y^k)}.$$

Покажем, что последовательность $\{\Lambda_1^k\}$ сходиться к λ_1.

Представим начальное приближение y^0 в виде разложения по базису из собственных векторов матрицы A. То есть $y^0 = \sum_{i=1}^{n} \alpha_i \xi_i$.

Из (2.4) следует, что $y^k = Ay^{k-1} = \ldots = A^k y^0$.

Тогда верно следующее

$$y^k = A^k \sum_{i=1}^{n} \alpha_i \xi_i = A^{k-1} \sum_{i=1}^{n} \alpha_i A\xi_i = A^{k-1} \sum_{i=1}^{n} \alpha_i \lambda_i \xi_i = \ldots = \sum_{i=1}^{n} \alpha_i \lambda_i^k \xi_i.$$

Вычислим два скалярных произведения :

$$(y^k, y^k) = (\sum_{i=1}^{n} \alpha_i \lambda_i^k \xi_i, \sum_{j=1}^{n} \alpha_j \lambda_j^k \xi_j) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \lambda_i^k \lambda_j^k (\xi_i, \xi_j) =$$

$$= \sum_{i=1}^{n} \alpha_i^2 \lambda_i^{2k} = \alpha_1^2 \lambda_1^{2k} + \sum_{i=2}^{n} \alpha_i^2 \lambda_i^{2k};$$

$$(y^{k+1}, y^k) = (\sum_{i=1}^{n} \alpha_i \lambda_i^{k+1} \xi_i, \sum_{j=1}^{n} \alpha_j \lambda_j^k \xi_j) = \ldots =$$

$$= \sum_{i=1}^{n} \alpha_i^2 \lambda_i^{2k+1} = \alpha_1^2 \lambda_1^{2k+1} + \sum_{i=2}^{n} \alpha_i^2 \lambda_i^{2k+1}.$$

Тогда, для элементов последовательности Λ_1^k, получаем

$$\Lambda_1^k = \frac{(y^{k+1}, y^k)}{(y^k, y^k)} = \frac{\alpha_1^2 \lambda_1^{2k+1} + \sum_{i=2}^{n} \alpha_i^2 \lambda_i^{2k+1}}{\alpha_1^2 \lambda_1^{2k} + \sum_{i=2}^{n} \alpha_i^2 \lambda_i^{2k}} =$$

$$= \frac{\alpha_1^2 \lambda_1^{2k+1} \left(1 + \sum_{i=2}^{n} \left(\frac{\alpha_i}{\alpha_1}\right)^2 \left(\frac{\lambda_i}{\lambda_1}\right)^{2k+1}\right)}{\alpha_1^2 \lambda_1^{2k} \left(1 + \sum_{i=2}^{n} \left(\frac{\alpha_i}{\alpha_1}\right)^2 \left(\frac{\lambda_i}{\lambda_1}\right)^{2k}\right)} =$$

$$= \lambda_1 \frac{1 + O\left(\left(\frac{\lambda_2}{\lambda_1}\right)^{2k+1}\right)}{1 + O\left(\left(\frac{\lambda_2}{\lambda_1}\right)^{2k}\right)} = \lambda_1 \left(1 + O\left(\left(\frac{\lambda_2}{\lambda_1}\right)^{2k}\right)\right) \longrightarrow \lambda_1$$

при $k \to \infty$.

Последовательность Λ_1^k сходится к λ_1 - искомому максимальному по модулю собственному значению матрицы A.

Рассмотрим последовательность векторов $\frac{y^k}{||y^k||}$. Верны следующие преобразования

$$\frac{y^k}{||y^k||} = \frac{\alpha_1\lambda_1^k\xi_1 + \sum_{i=2}^{n}\alpha_i\lambda_i^k\xi_i}{\sqrt{\alpha_1^2\lambda_1^{2k} + \sum_{i=2}^{n}\alpha_i^2\lambda_i^{2k}}} = \frac{\alpha_1\lambda_1^k\xi_1 + \sum_{i=2}^{n}\alpha_i\lambda_i^k\xi_i}{|\alpha_1||\lambda_1|^k\left(1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{2k}\right)\right)} =$$

$$= \pm\xi_1 + \sum_{i=2}^{n}\frac{\alpha_i}{|\alpha_1|}O\left(\left(\frac{\lambda_i}{\lambda_1}\right)^k\right)\xi_i.$$

То есть, вектор $\frac{y^k}{||y^k||}$ с ростом итерационного индекса k приближается к направлению собственного вектора ξ_1 .

2.2.1 Поиск максимального и минимального собственного значения

Алгоритм вычисления максимального по модулю собственного значения матрицы $A = A^T$ можно использовать для поиска максимального и минимального собственного значения матрицы A. Обозначим через $\Lambda(A)$ максимальное по модулю собственное значение матрицы A, вычисленное с использованием рассмотренного алгоритма. Сформируем вспомогательную матрицу $D = A - \Lambda(A)E$. Собственные числа матрицы D равны $\lambda_i(D) = \lambda_i(A) - \Lambda(A)$, $i = 1, \ldots, n$.

Пусть вычисленное $\Lambda(A) < 0$. Тогда $\lambda_{\min}(A) = \Lambda(A)$ и все собственные значения матрицы D удовлетворяют условию $\lambda_i(D) \geq 0$. Следовательно, вычислив максимальное по модулю собственное значение матрицы D, получим $\Lambda(D) = \lambda_{\max}(A) - \Lambda(A)$. Отсюда $\lambda_{\max}(A) = \Lambda(D) + \Lambda(A)$.

Пусть выполнено условие $\Lambda(A) > 0$. Тогда $\lambda_{\max}(A) = \Lambda(A)$ и собственные значения матрицы D удовлетворяют условию $\lambda_i(D) \leq 0$. Следовательно, для

максимального по модулю собственного значения матрицы D верно равенство $\Lambda(D) = \lambda_{\min}(A) - \Lambda(A)$. Отсюда $\lambda_{\min}(A) = \Lambda(D) + \Lambda(A)$.

2.2.2 Поиск собственного значения ближайшего к заданному числу

Рассмотрим задачу о поиске такого λ, что

$$|\lambda - a| = \min_i |\lambda_i(A) - a|,$$

где a — заданное число. Построим вспомогательную матрицу $D = E - c(A - aE)^2$, в которой числовой параметр $c > 0$ подбирается так, чтобы выполнялось неравенство $\Lambda(-c(A - aE)^2) > -1$. Тогда справедливо уравнение $\Lambda(D) = 1 - c(\lambda - a)^2$, решением которого является искомое собственное значение λ.

2.3 Метод обратных итераций

Пусть для невырожденной матрицы A известно приближенное значение $\overline{\lambda}$ собственного числа $\lambda(A)$. Требуется найти собственный вектор матрицы A, соответствующий собственному значению $\lambda(A)$.

Рассмотрим линейную систему уравнений

$$(A - \overline{\lambda}E)y = b, \tag{2.5}$$

где произвольный вектор $b \neq 0$.

Покажем, что решение этой системы будет приближенно равняться искомому собственному вектору матрицы A. Пусть матрица A такова, что ее собственные вектора $\{\xi_i\}$ образуют базис. Пусть $b = \sum_j \beta_j \xi_j$ и $y = \sum_j \alpha_j \xi_j$ — разложения вектора b и искомого вектора y по этому базису. Подставим данные разложения в (2.5)

$$\begin{gathered}(A - \overline{\lambda}E)\sum_j \alpha_j \xi_j = \sum_j \beta_j \xi_j, \\ \sum_j (\alpha_j \lambda_j - \alpha_j \overline{\lambda})\xi_j = \sum_j \beta_j \xi_j, \\ \sum_j [\alpha_j(\lambda_j - \overline{\lambda}) - \beta_j]\xi_j = 0.\end{gathered}$$

Коэффициенты, равные нулю, обращают линейную комбинацию базисных векторов в нуль. Следовательно, получаем выражение для α_j :

$$\alpha_j = \frac{\beta_j}{\lambda_j - \overline{\lambda}}.$$

То есть, в разложении вектора y по базису из собственных векторов матрицы A коэффициент α_j при базисном векторе ξ_j будет превосходить по абсолютной величине все другие коэффициенты, если λ_j близко к значению $\overline{\lambda}$. Поэтому говорят, что вектор y близок к собственному вектору ξ_j по направлению.

Для усиления этого эффекта строится последовательность векторов y^k по следующему правилу

$$(A - \overline{\lambda}E)y^{k+1} = y^k, \quad k = 0, 1, \ldots,$$

где $y^0 = b$. Данный итерационный процесс называется методом обратных итераций.

Глава 3

Численные методы решения нелинейных уравнений

Пусть функция $f(x)$ определена и непрерывна на отрезке $[a;\ b]$. Требуется найти корни уравнения $f(x) = 0$ на заданном отрезке.

Решение задачи можно разбить на два этапа. На первом этапе проводится разделение корней. То есть, выделяются участки области определения, содержащие только один корень. Затем на каждом из этих участков, с использованием итерационного процесса, проводится уточнение значения искомого корня.

3.1 Методы разделения корней

Один из способов разделения или локализации корней состоит в следующем. Отрезок $[a;\ b]$ произвольным образом разбивается на N частей $[x_j;\ x_{j+1}]$:

$$a = x_0 < x_1 < x_2 < \ldots < x_N = b.$$

Вычисляются значения функции $f(x)$ в точках x_j : $f(x_j) = f_j$, $j = 0, 1, \ldots N$. Отбираются те отрезки $[x_j;\ x_{j+1}]$, для которых выполняется условие

$$f_j \cdot f_{j+1} < 0,$$

означающее, что непрерывная функция $f(x)$ имеет на отрезке $[x_j;\ x_{j+1}]$ корень.

Затем, каждый из выделенных отрезков, вновь разбивается на части, для которых повторяется предыдущая процедура. В результате удается найти участки достаточно малой длины, содержащие искомые корни.

Вариантом предыдущего способа является метод бисекции. Пусть $f(a) \cdot f(b) < 0$ — это означает, что внутри $[a;\ b]$ есть корень уравнения $f(x) = 0$.

Выберем $x_0 = \frac{a+b}{2}$ — середина отрезка $[a;\, b]$. Если x_0 не является корнем, то либо $f(a)\cdot f(x_0) < 0$, либо $f(x_0)\cdot f(b) < 0$. Пусть выполняется неравенство для отрезка $[a;\, x_0]$. Тогда выбираем $x_1 = \frac{a+x_0}{2}$ и проверяем выполнение неравества $f(a)\cdot f(x_1) < 0$ или $f(x_1)\cdot f(x_0) < 0$. Процесс деления отрезков пополам заканчивается, когда длина очередного отрезка станет меньше заданной величины. Данный процесс сходится к одному из корней функции $f(x)$на отрезке $[a;\, b]$.

3.2 Примеры итерационных методов вычисления корней

3.2.1 Метод простой итерации

Рассмотрим уравнение $f(x) = 0$. Пусть $x^* \in [a;\, b]$ — корень уравнения. Пусть вспомогательная непрерывная функция $\tau(x) > 0$ на отрезке $[a;\, b]$. Тогда

$$f(x) = 0 \iff -\tau(x)f(x) = 0 \iff x - \tau(x)f(x) = x.$$

Введём обозначение $S(x) = x - \tau(x)f(x)$. Итак

$$f(x) = 0 \iff S(x) = x.$$

Построим числовую последовательность $\{x^k\}$ по следующему правилу

$$x^{k+1} = S(x^k), \quad k = 0, 1, \ldots,$$

начальное приближение x^0 — задано.

Если предел последовательности $\{x^k\}$ существует и $\lim\limits_{k\to\infty} x^k = x^*$ то, в силу непрерывности $S(x)$, верно равенство $x^* = S(x^*)$ и, следовательно, x^* является корнем исходного уравнения.

На сходимость последовательности $\{x^k\}$ влияет функция $S(x)$, которая содержит параметр $\tau(x)$.

Проведём некоторые геометрические построения. Искомому корню x^* соответствует точка пересечения линий $y(x) = x$ и $y(x) = S(x)$. Пусть функция $S(x)$ является дифференцируемой функцией и в некоторой окрестности искомого корня x^* выполнено условие $0 < S'(x) < 1$. Тогда в данной окрестности линии $y(x) = x$ и $y(x) = S(x)$ располагаются, например, так, как изображено на рисунке (Рис.3.1).

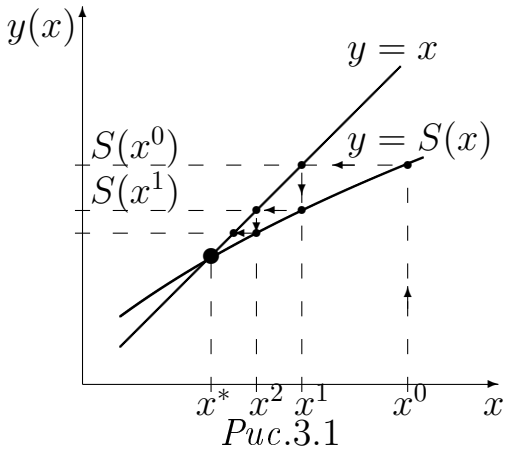

Рис.3.1

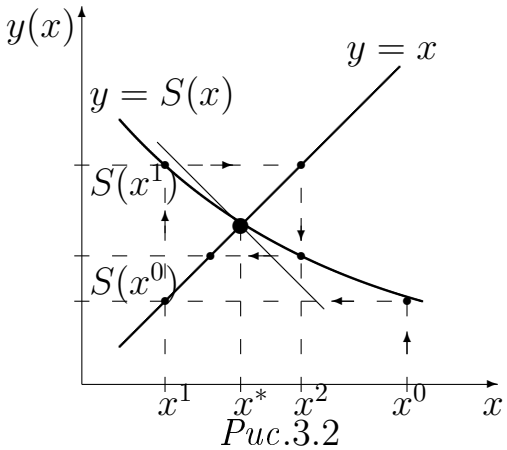

Рис.3.2

Видно, что выбрав начальное приближение x^0 вблизи искомого корня x^*, получаем последовательность $\{x^k\}$, сходящуюся к корню.

Пусть в окрестности искомого корня x^* выполнено условие $-1 < S'(x) < 0$. Тогда, в окрестности корня линии $y(x) = x$ и $y(x) = S(x)$ располагаются так, как изображено на Рис.3.2. Видно, что и в этом случае выбрав начальное приближение x^0 вблизи корня, получаем сходящуюся к этому корню последовательность $\{x^k\}$.

В других ситуациях последовательность $\{x^k\}$ оказывается расходящейся. Пусть в окрестности корня x^* выполнено условие $1 < S'(x)$. Тогда в данной окрестности линии $y(x) = x$ и $y(x) = S(x)$ располагаются так, как показано на Рис.3.3.

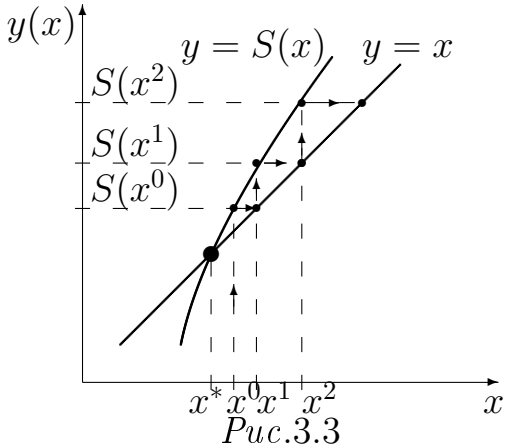

Рис.3.3

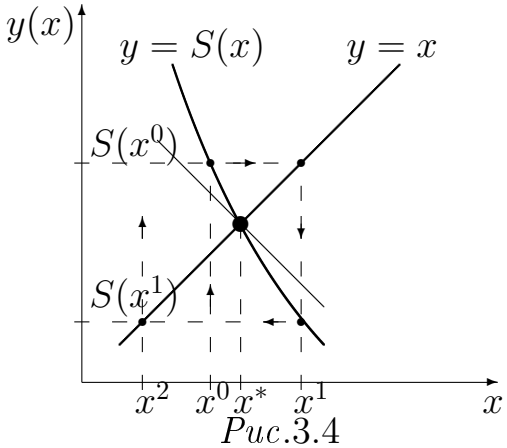

Рис.3.4

То есть, для любого начального приближения x^0, получаем расходящуюся последовательность итерационных приближений $\{x^k\}$.

Пусть в окрестности корня x^* выполнено условие $S'(x) < -1$. Тогда, в окрестности корня линии $y(x) = x$ и $y(x) = S(x)$ располагаются так, как изображено на Рис.3.4. В этом случае, выбрав начальное приближение x^0 вблизи корня, получаем расходящуюся итерационную последовательность $\{x^k\}$.

Можно сделать предположение, что достаточным условием сходимости метода простой итерации является выполнение неравенства $|S'(x)| < 1$ в окрестности корня x^*.

3.2.2 Метод Ньютона

Пусть x^* — корень функции $f(x)$, то есть $f(x^*) = 0$. Пусть $f'(x)$ существует, непрерывна и отлична от нуля в некоторой окрестности корня. Запишем исходное уравнение в виде $f(x^k + (x^* - x^k)) = 0$, где x^k — заданное итерационное приближение для корня x^*. Применим к данному выражению формулу Лагранжа

$$f(x^k) + f'(\overline{x})(x^* - x^k) = 0,\ \overline{x} \in [x^*;\ x^k].$$

Заменим в этом соотношении $\overline{x}$ на x^k, а x^* на x^{k+1} — следующее итерационное приближение, которое удовлетворяет уравнению $f(x^k) + f'(x^k)(x^{k+1} - x^k) = 0$. Тогда, элементы числовой последовательности итерационных приближений, построенных по методу Ньютона, вычисляются по формуле

$$x^{k+1} = x^k - \frac{f(x^k)}{f'(x^k)}, \quad k = 0, 1, \ldots, \tag{3.1}$$

начальное приближение x^0 — задано.

Метод Ньютона называют также методом касательных. Поясняет такое название следующий рисунок (Рис.3.5). На Рис.3.5 в некоторой окрестности искомого корня x^* представлен график функции $f(x)$.

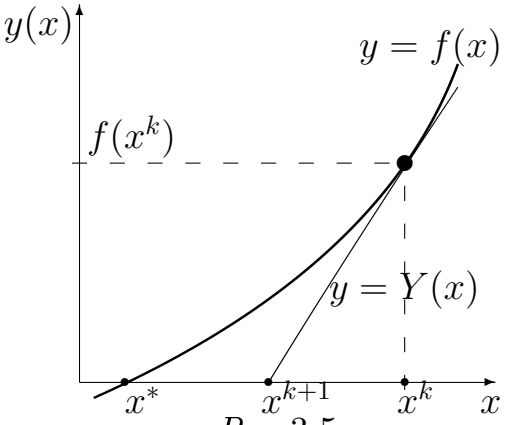

Рис.3.5

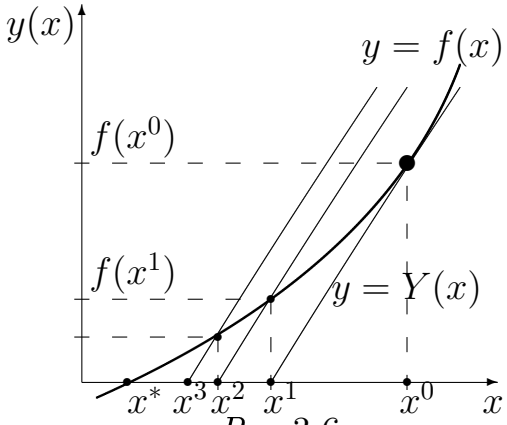

Рис.3.6

Пусть x^k — заданное итерационное приближение к корню x^*. Через точку плоскости с координатами $x = x^k$ и $y = f(x^k)$ проведена прямая $Y(x)$, являющаяся касательной к графику функции $f(x)$ в точке $x = x^k$. Уравнение прямой $Y(x)$ имеет вид

$$Y(x) = f'(x^k)x + \Big(f(x^k) - f'(x^k)x^k\Big) = (x - x^k)f'(x^k) + f(x^k).$$

Точку пересечения прямой $Y(x)$ с осью x примем за x^{k+1} итерационное приближение к искомому корню x^*. Тогда условие $Y(x^{k+1}) = 0$ приводит к уравнению относительно x^{k+1} вида $(x^{k+1} - x^k)f'(x^k) + f(x^k) = 0$, из которого следует формула (3.1) для x^{k+1}.

Замечание. 1. Метод Ньютона формально можно считать методом простой итерации с функцией $S(x)$ специального вида

$$S(x) = x - \tau(x)f(x) = \left\{\tau(x) = \frac{1}{f'(x)}\right\} = x - \frac{f(x)}{f'(x)}.$$

Выполнение в окрестности корня x^* неравенства $|S'(x)| < 1$ является достаточным условием для сходимости последовательности итерационных приближений к корню x^*. В случае метода Ньютона достаточное условие сходимости принимает вид

$$|S'(x)| = \left|1 - \frac{(f'(x))^2 - f(x)f''(x)}{(f'(x))^2}\right| = \left|\frac{f(x)f''(x)}{(f'(x))^2}\right| < 1.$$

В малой окрестности корня x^* значение функции $f(x)$ близко к нулю и условие сходимости может выполняться.

2. Особенностью метода Ньютона является квадратичная скорость сходимости итерационных приближений к корню x^*. То есть, $\left|x^{k+1} - x^*\right| = O\left(\left|x^k - x^*\right|^2\right)$.

3.2.3 Модифицированный метод Ньютона

В случае, когда вычисление на каждой итерации производной $f'(x^k)$ является трудоемкой операцией, можно использовать модифицированный метод Ньютона, расчётная формула которого имеет вид

$$x^{k+1} = x^k - \frac{f(x^k)}{f'(x^0)}, \quad k = 0, 1, \ldots,$$

где x^0 — заданное начальное приближение.

Модифицированный метод Ньютона можно назвать методом одной касательной (Рис.3.6). Через точку $x = x^0$ и $y = f(x^0)$ проводится прямая, касательная к графику функции $y = f(x)$. Точку пересечения этой прямой с осью x принимают за x^1 итерационное приближение. Следующими итерационными приближениями являются точки пересечения параллельных прямых $Y(x) = (x - x^k)f'(x^0) + f(x^k), \quad k = 0, 1, \ldots,$ с осью x.

Замечание. Модифицированный метод Ньютона имеет линейную скорость сходимости, то есть $\left|x^{k+1} - x^*\right| = O(\left|x^k - x^*\right|)$.

3.2.4 Метод секущих

В случае, когда нет возможности вычислить производную $f'(x^k)$, заменим ее в формуле (3.1) на разностное отношение $f'(x^k) \approx \dfrac{f(x^k) - f(x^{k-1})}{x^k - x^{k-1}}$.

Тогда получим формулу, определяющую метод секущих

$$x^{k+1} = x^k - \frac{x^k - x^{k-1}}{f(x^k) - f(x^{k-1})} f(x^k), \quad k = 1, 2, \ldots . \tag{3.2}$$

Для начала расчёта по формуле (3.2) необходимо задать x^0 и x^1 — два начальных приближения.

Используем Рис.3.7 для пояснения названия метода.

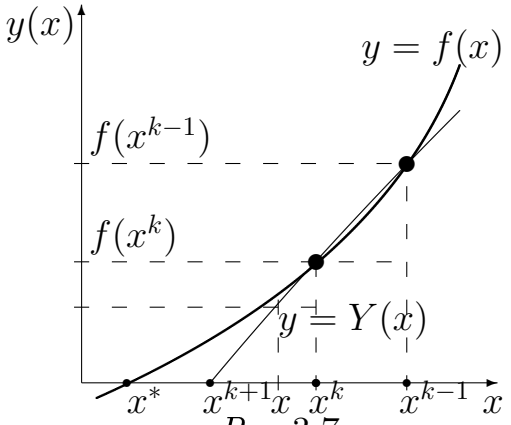

*Рис.*3.7

Уравнение прямой $Y(x)$, проходящей через точки $(x^{k-1},\ f(x^{k-1}))$ и $(x^k,\ f(x^k))$ имеет вид

$$\frac{Y(x) - f(x^k)}{x - x^k} = \frac{f(x^k) - f(x^{k-1})}{x^k - x^{k-1}}.$$

Прямую $Y(x)$ можно считать секущей для графика $y = f(x)$. Точку пересечения прямой $Y(x)$ с осью x примем за x^{k+1} итерационное приближение к корню x^*. Итак, пусть в предыдущем соотношении $Y(x) = 0$ и $x = x^{k+1}$. Тогда, для вычисления x^{k+1} получим формулу (3.2).

Замечание. 1. Метод секущих имеет линейную скорость сходимости, то есть $\left|x^{k+1} - x^*\right| = O(\left|x^k - x^*\right|)$.

2. Метод секущих применим, например, в следующей ситуации. Пусть есть компьютерная программа $P(x)$, вычисляющая по заданному входному параметру x некоторую выходную величину $y = P(x)$. Представляет интерес получение конкретного выходного значения $y = y^*$. Необходимо определить значение параметра $x = x^*$, при котором $P(x^*) = y^*$.

Рассмотрим функцию $f(x) = y^* - P(x)$. Корень этой функции является искомым параметром x^*. Вычислить корень можно с помощью метода секущих.

3.3 Сходимость метода простой итерации

В методе простой итерации элементы последовательности итерационных приближений $\{x^k\}$ вычисляются по формуле $x^{k+1} = S(x^k), \quad k = 0, 1, \ldots$. Начальное приближение x^0 — задано.

Введем обозначение $U_r(a) = \{x : |x - a| \leqslant r\}$.

Верно следующее утверждение о сходимости метода простой итерации.

Теорема 3.1. Пусть для функции $S(x)$ на множестве $U_r(a)$ выполняется неравенство $|S(x') - S(x'')| \leqslant q|x' - x''|$ для любых x', $x'' \in U_r(a)$ с константой $q \in (0; 1)$. Параметр a выбран так, что $|S(a) - a| \leqslant (1 - q)r$.

Тогда уравнение $x = S(x)$ имеет на множестве $U_r(a)$ единственный корень x^* и последовательность $\{x^k\}$ сходится к x^* при любом начальном приближении $x^0 \in U_r(a)$. Для погрешности итерационного приближения справедлива оценка $|x^k - x^*| \leqslant q^k|x^0 - x^*|$.

Доказательство. Выберем начальное приближение $x^0 \in U_r(a)$. Пусть $x^j \in U_r(a)$. Покажем, что следующее итерационное приближение $x^{j+1} \in U_r(a)$.

Верны следующие преобразования:

$$\begin{aligned} |x^{j+1} - a| = |S(x^j) - a| = |S(x^j) - S(a) + S(a) - a| \leqslant \\ \leqslant |S(x^j) - S(a)| + |S(a) - a| \leqslant q|x^j - a| + (1 - q)r \leqslant \\ \leqslant qr + (1 - q)r = r. \end{aligned}$$

Оценим разность двух соседних итерационных приближений:

$$\begin{aligned} |x^{j+1} - x^j| = |S(x^j) - S(x^{j-1})| \leqslant q|x^j - x^{j-1}| = \\ = q|S(x^{j-1}) - S(x^{j-2})| \leqslant q^2|x^{j-1} - x^{j-2}| \leqslant \ldots \leqslant q^j|x^1 - x^0|. \end{aligned}$$

Покажем, что последовательность $\{x^k\}$ имеет предел. Используем критерий Коши сходимости числовой последовательности:

$$\begin{aligned} |x^{k+p} - x^k| = \left|\sum_{j=1}^{p} \left(x^{k+j} - x^{k+j-1}\right)\right| \leqslant \sum_{j=1}^{p} \left|x^{k+j} - x^{k+j-1}\right| \leqslant \\ \leqslant \sum_{j=1}^{p} q^{k+j-1}|x^1 - x^0| = q^k|x^1 - x^0| \sum_{j=1}^{p} q^{j-1} < \\ < q^k|x^1 - x^0| \sum_{j=1}^{\infty} q^{j-1} = \frac{q^k}{1 - q}|x^1 - x^0|. \end{aligned}$$

Для любого $\varepsilon > 0$ выражение $\dfrac{q^k}{1-q}|x^1 - x^0|$ будет меньше этого ε если

$$k > k_0(\varepsilon) = \left[\frac{\ln \frac{|S(x^0)-x^0|}{\varepsilon(1-q)}}{\ln(1/q)}\right].$$

Таким образом, числовая последовательность $\{x^k\}$ сходится при $k \to \infty$ к некоторому $x^* \in U_r(a)$. Покажем, что x^* является корнем уравнения $x = S(x)$.

Верно неравенство $|S(x^k) - S(x^*)| \leqslant q|x^k - x^*|$, из которого следует, что $S(x^k) \xrightarrow[k\to\infty]{} S(x^*)$. Перейдём в соотношении $x^{k+1} = S(x^k)$ к пределу при $k \to \infty$. В результате получим, что $x^* = S(x^*)$.

Покажем единственность корня x^*. Пусть существуют два различных корня x^* и $\overline{x^*}$. Тогда $|x^* - \overline{x^*}| = |S(x^*) - S(\overline{x^*})| \leqslant q|x^* - \overline{x^*}|$, что противоречит неравенству $q < 1$. Следовательно, $x^* = \overline{x^*}$.

Для погрешности итерационного приближения верна следующая оценка

$$|x^k - x^*| = |S(x^{k-1}) - S(x^*)| \leqslant q|x^{k-1} - x^*| = \\ = q|S(x^{k-2}) - S(x^*)| \leqslant q^2|x^{k-2} - x^*| \leqslant \cdots \leqslant q^k|x^0 - x^*|.$$

Теорема доказана. □

Замечание. 1. Была получена оценка $|x^{k+p} - x^k| \leqslant \dfrac{q^k}{1-q}|x^1 - x^0|$, верная для любого натурального p. Переходя к пределу при $p \to \infty$, имеем $|x^* - x^k| \leqslant \dfrac{q^k}{1-q}|S(x^0) - x^0|$.

Потребуем, чтобы $\dfrac{q^k}{1-q}|S(x^0) - x^0| \leqslant \varepsilon$. Тогда, x^k будет отличаться от x^* не более чем на ε. То есть, $|x^* - x^k| \leqslant \varepsilon$ при

$$k \geq k(\varepsilon) = \left[\frac{\ln \frac{|S(x^0)-x^0|}{\varepsilon(1-q)}}{\ln(1/q)}\right].$$

2. Пусть у функции $S(x)$ существует производная, которая удовлетворяет неравенству $|S'(x)| \leq q$ для любого $x \in U_r(a)$. Тогда $|S(x') - S(x'')| = |S'(\xi)||x' - x''| \leqslant q|x' - x''|$ для любых x', $x'' \in U_r(a)$. То есть, выполняется одно из предположений теоремы о сходимости метода простой итерации.

3.4 Метод Эйткена

Пусть x^{k+1}, x^k и x^{k-1} три последовательных итерационных приближения, вычисленных с помощью итерационного метода, имеющего линейную скорость сходимости. Например, с испопользованием метода простой итерации. Расположение этих итерационных приближений в окрестности корня x^* в случае выполнения неравенства $0 < S'(x) < 1$ для метода простой итерации иллюстрирует Рис.3.8.

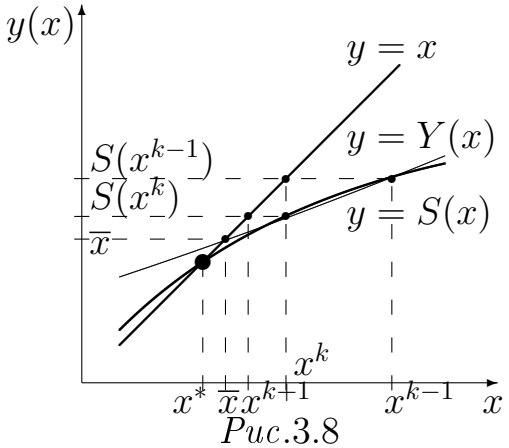

*Рис.*3.8

Проведём через две точки $(x^{k-1},\, S(x^{k-1}))$ и $(x^k,\, S(x^k))$ прямую $y = Y(x)$. Уравнение прямой $Y(x)$ можно записать, например, в виде

$$\frac{S(x^{k-1}) - S(x^k)}{x^{k-1} - x^k} = \frac{S(x^{k-1}) - Y(x)}{x^{k-1} - x}.$$

Точка пересечения прямых $y = Y(x)$ и $y = x$ имеет координаты $x = \overline{x}$, $y = \overline{x}$. Подставляя в уравнение прямой $Y(x)$ значения $x = \overline{x}$, $Y(x) = \overline{x}$ и $S(x^{k-1}) = x^k$, $S(x^k) = x^{k+1}$ имеем уравнение для определения $\overline{x}$

$$\frac{x^k - x^{k+1}}{x^{k-1} - x^k} = \frac{x^k - \overline{x}}{x^{k-1} - \overline{x}}.$$

Отсюда получаем, что

$$\overline{x} = \frac{x^{k+1}x^{k-1} - (x^k)^2}{x^{k+1} - 2x^k + x^{k-1}}.$$

Значение $\overline{x}$ ближе к искомому корню x^*, чем x^{k+1}. Примем вычисленное $\overline{x}$ в качестве нового начального приближения для используемого итерационного метода. Периодическое повторение данной процедуры называется методом Эйткена ускорения сходимости линейно сходящихся итерационных методов.

3.5 Сходимость метода Ньютона

В методе Ньютона итерационные приближения вычисляются по формуле

$$x^{k+1} = x^k - \frac{f(x^k)}{f'(x^k)},\ k = 0, 1, \dots,$$

x^0 —задано. Для последовательности итерационных приближений $\{x^k\}$ верно следующее утверждение о сходимости.

Теорема 3.2. Пусть x^* — простой вещественный корень уравнения $f(x) = 0$. Функция $f(x)$ — дважды дифференцируема в некоторой окрестности $U_r(x^*)$ и $f'(x) \neq 0$ при $x \in U_r(x^*)$. Пусть $0 < m = \inf_{x \in U_r(x^*)} |f'(x)|$ и $M = \sup_{x \in U_r(x^*)} |f''(x)|$, а начальное приближение $x^0 \in U_r(x^*)$ выбрано так, что комбинация

$$\frac{M|x^0 - x^*|}{2m} = q < 1.$$

Тогда итерационная последовательность

$$x^{k+1} = x^k - \frac{f(x^k)}{f'(x^k)},\ k = 0,\ 1,\ \dots$$

сходиться к x^* и для погрешности итерационного приближения верна оценка

$$|x^k - x^*| \leqslant q^{2^k - 1}|x^0 - x^*|. \tag{3.3}$$

Доказательство. Погрешность $(k+1)$-ого итерационного приближения запишем в виде

$$x^{k+1} - x^* = x^k - \frac{f(x^k)}{f'(x^k)} - x^* = \frac{(x^k - x^*)f'(x^k) - f(x^k)}{f'(x^k)} = \frac{F(x^k)}{f'(x^k)},$$

где $F(x) = (x - x^*)f'(x) - f(x)$.

Так как $F(x^*) = f(x^*) = 0$, то

$$F(x^k) = \int\limits_{x^*}^{x^k} F'(\xi)d\xi = \int\limits_{x^*}^{x^k} (\xi - x^*)f''(\xi)d\xi = f''(\xi_k)\frac{(x^k - x^*)^2}{2},$$

где $\xi_k \in [x^*;\ x^k] \in U_r(x^*)$.

То есть, погрешность представима в виде

$$x^{k+1} - x^* = f''(\xi_k)\frac{(x^k - x^*)^2}{2f'(x^k)}. \tag{3.4}$$

Докажем оценку для погрешности итерационных приближений из условий теоремы по индукции.

База индукции. Рассмотрим соотношение (3.4) при $k = 0$

$$|x^1 - x^*| = |f''(\xi_0)|\frac{|x^0 - x^*|^2}{2|f'(x^0)|} \leqslant \frac{M|x^0 - x^*|}{2m}|x^0 - x^*| = q|x^0 - x^*|.$$

Таким образом, неравенство (3.3), в котором $k = 1$, выполняется и база индукции верна.

Предположение индукции. Пусть оценка (3.3) выполняется для некоторого k, то есть $|x^k - x^*| \leqslant q^{2^k-1}|x^0 - x^*|$.

Индуктивный переход. Покажем, что оценка (3.3) выполняется и для $(k+1)$-ого итерационного приближения. Из (3.4) следует, что

$$|x^{k+1} - x^*| = |f''(\xi^k)|\frac{|x^k - x^*|^2}{2|f'(x^k)|} \leqslant \frac{M}{2m}|x^k - x^*|^2 \leqslant$$
$$\leqslant \frac{M}{2m}(q^{2^k-1})^2|x^0 - x^*|^2 = \frac{M|x^0 - x^*|}{2m}q^{2^{k+1}-2}|x^0 - x^*| =$$
$$= \left\{\frac{M|x^0 - x^*|}{2m} = q\right\} = q^{2^{k+1}-1}|x^0 - x^*|.$$

Оценка (3.3) верна. Выполнив в (3.3) предельный переход при $k \to \infty$ получим, что, так как $q < 1$, правая часть стремится к нулю и, следовательно, последовательность $\{x^k\}$ сходится к x^*.

Теорема доказана. □

Замечание. В условиях теоремы предполагалось выполненным неравенство $\dfrac{M|x^0 - x^*|}{2m} < 1$. Добиться этого можно следующим образом.

Запишем $f(x^0)$ в виде $f(x^0) = f(x^0) - f(x^*) = f'(\overline{x})(x^0 - x^*)$.

Тогда $|x^0 - x^*| = \dfrac{|f(x^0)|}{|f'(\overline{x})|} \leqslant \dfrac{|f(x^0)|}{m}$.

Следовательно $\dfrac{M|x^0 - x^*|}{2m} \leqslant \dfrac{M|f(x^0)|}{2m^2}$.

Таким образом, зная m и M, нужно подобрать $x^0 \in U_r(x^*)$ так, чтобы было верно неравенство $\dfrac{M|f(x^0)|}{2m^2} < 1$.

3.6 Решение систем нелинейных уравнений

Пусть имеется n нелинейных уравнений

$$f_i(x_1,\, x_2,\, \ldots,\, x_n) = 0, \quad i = \overline{1,\, n}.$$

Поиск корней системы нелинейных уравнений осуществляется в два этапа. На первом этапе проводят локализацию корней, а затем, используя итерационные методы, вычисляют корни с требуемой точностью.

Метод Ньютона

Пусть $x^* = (x_1^*,\, x_2^*,\, \ldots,\, x_n^*)^T$ корень системы нелинейных уравнений

$$f_i(x_1^*,\, x_2^*,\, \ldots,\, x_n^*) = 0, \quad i = \overline{1,\, n}. \tag{3.5}$$

Получим расчетные формулы метода Ньютона для решения системы нелинейных уравнений. Пусть известно k-ое итерационное приближение $x^k = (x_1^k,\, x_2^k,\, \ldots,\, x_n^k)^T$ для искомого корня $x^* = (x_1^*,\, x_2^*,\, \ldots,\, x_n^*)^T$. Запишем аргументы функций f_i из (3.5) в виде

$$f_i(x_1^k + (x_1^* - x_1^k),\, x_2^k + (x_2^* - x_2^k),\, \ldots,\, x_n^k + (x_n^* - x_n^k)) = 0, \quad i = \overline{1,\, n}.$$

Пусть у функций $f_i(x_1,\, x_2,\, \ldots,\, x_n)$ существуют непрерывные частные производные первого порядка по всем аргументам. Тогда, используя формулу Тейлора с остаточным членом в форме Лагранжа, получим

$$f_i(x_1^k,\, x_2^k,\, \ldots,\, x_n^k) + \sum_{l=1}^{n} \frac{\partial f_i(\xi_1^k,\, \xi_2^k,\, \ldots,\, \xi_n^k)}{\partial x_l}(x_l^* - x_l^k) = 0, \quad i = \overline{1,\, n}.$$

В этих соотношениях неизвестны значения $\xi^k = (\xi_1^k,\, \xi_2^k,\, \ldots,\, \xi_n^k)^T$ и $x^* = (x_1^*,\, x_2^*,\, \ldots,\, x_n^*)^T$. Заменим ξ^k на x^k, а x^* на x^{k+1}, которое примем за следующее $(k+1)$-ое итерационное приближение к корню x^*. В результате получим

$$f_i(x_1^k,\, x_2^k,\, \ldots,\, x_n^k) + \sum_{l=1}^{n} \frac{\partial f_i(x_1^k,\, x_2^k,\, \ldots,\, x_n^k)}{\partial x_l}(x_l^{k+1} - x_l^k) = 0, \quad i = \overline{1,\, n}. \tag{3.6}$$

Система (3.6) является линейной системой алгебраических уравнений относительно $x^{k+1} = (x_1^{k+1},\, x_2^{k+1},\, \ldots,\, x_n^{k+1})^T$. Введём обозначение $\Delta x_l^k = x_l^{k+1} - x_l^k$ и запишем (3.6) в виде

$$\sum_{l=1}^{n} \frac{\partial f_i(x_1^k,\, x_2^k,\, \ldots,\, x_n^k)}{\partial x_l}\Delta x_l^k = -f_i(x_1^k,\, x_2^k,\, \ldots,\, x_n^k), \quad i = \overline{1,\, n}.$$

Решив эту систему, находим вектор $\Delta x^k = (\Delta x_1^k, \Delta x_2^k, \ldots, \Delta x_n^k)^T$. Затем вычисляем следующее итерационное приближение $x^{k+1} = x^k + \Delta x^k$.

Замечание. Система линейных уравнений (3.6) имеет решение если в некоторой окрестности искомого корня x^* определитель матрицы, составленной из частных производных функций $f_i(x_1, x_2, \ldots, x_n)$ по всем их аргументам, отличен от нуля.

Последовательность векторов x^k быстро сходится к корню x^*, так как метод Ньютона имеет квадратичную скорость сходимости.

3.7 Примеры

3.7.1 Решение нелинейного уравнения

Пусть $f(x) = x^3 - x - 1$. Найдем корень уравнения $x^3 - x - 1 = 0$, расположенный на отрезке $[-2; 3]$. Проведем локализацию корня. Выберем на отрезке $[-2; 3]$ набор точек x_i и вычислим значения функции в точках x_i:

x_i	-2	-1	0	1	2	3
$f(x_i)$	-7	-1	-1	-1	5	23

Так как $f(1)f(2) < 0$, то корень уравнения находится на отрезке $[1; 2]$. Искомый корень $x^* \approx 1.324717957$.

1. Метод простой итерации.

Используем метод простой итерации для вычисления корня.

1.1

Преобразуем уравнение $f(x) = 0$ к виду $x = S(x)$, где $S(x) = x - \tau(x)f(x)$. Выберем $\tau(x) = \tau > 0$. Тогда $S(x) = x - \tau(x^3 - x - 1)$.

Достаточным условием сходимости последовательности итерационных приближений является выполнение неравенства $|S'(x)| < 1$ при $x \in [1; 2]$. Это неравенство приводит к ограничениям

$$|S'(x)| < 1 \iff |1 - \tau(3x^2 - 1)| < 1 \iff -1 < 1 - \tau(3x^2 - 1) < 1.$$

При $x \in [1; 2]$ и $\tau > 0$ правое неравенство верно всегда. Левое неравенство принимает вид $\tau < \dfrac{2}{3x^2 - 1}$. Знаменатель дроби достигает максимума при $x = 2$. Следовательно, τ должно удовлетворять условию $\tau < \dfrac{2}{3 \cdot 2^2 - 1} = \dfrac{2}{11}$.

Пусть $\tau = \dfrac{1}{11}$. В этом случае $S(x) = x - \dfrac{x^3 - x - 1}{11}$ и элементы последовательности итерационных приближений вычисляются по формуле

$$x^{k+1} = x^k - \frac{(x^k)^3 - x^k - 1}{11} \quad k = 0,\, 1,\, \ldots\,.$$

Результаты расчётов приведены в таблице

k	x^k	$\lvert x^k - x^*\rvert$
0	1.1	0.225
1	1.16991	0.155
2	1.22161	0.103
3	1.25784	0.067
4	1.28218	0.043
5	1.29802	0.027

Характер изменения величин итерационных приближений определяется следующим фактом. Погрешность итерационных приближений, вычисленных с использованием метода простой итерации, удовлетворяет условию $|x^{k+1} - x^*| \le q|x^k - x^*|$, в котором константа

$$q = \max_{x\in[1;\,2]} |S'(x)| = \max_{x\in[1;\,2]} \left|1 - \frac{3x^2 - 1}{11}\right| = \left.\frac{12 - 3x^2}{11}\right|_{x=1} = \frac{9}{11} \approx 0.8.$$

Значение $q \approx 0.8$ близко к единице и погрешности итерационных приближений уменьшаются медленно.

1.2

Приводить уравнение $f(x) = 0$ к виду $x = S(x)$ можно с учетом вида функции $f(x) = x^3 - x - 1$. Например,

$$x^3 - x - 1 = 0 \implies x = \sqrt[3]{x + 1}.$$

Тогда $S(x) = \sqrt[3]{x + 1}$. В этом случае скорость сходимости итерационных приближений x^k к корню x^* будет выше, так как при $x \in [1;\, 2]$

$$S'(x) = \frac{1}{3(x+1)^{\frac{2}{3}}} \le \left.\frac{1}{3(x+1)^{\frac{2}{3}}}\right|_{x=1} \approx 0.2 = q.$$

1.3

Для функции $f(x) = x^3 - x - 1$ возможно преобразование вида

$$x^3 - x - 1 = 0 \iff x = x^3 - 1.$$

То есть, $S(x) = x^3 - 1$ и $S'(x) = 3x^2 > 1$ при $x \in [1; 2]$. В этом случае последовательность итерационных приближений x^k будет расходящейся.

2. Метод Эйткена.

Воспользуемся методом Эйткена для ускорения сходимости линейно сходящихся итерационных процессов. В соответствии с методом Эйткена, по трём последовательным итерационным приближениям x^{k+1}, x^k и x^{k-1} вычисляется значение комбинации $\overline{x} = \dfrac{x^{k+1}x^{k-1} - (x^k)^2}{x^{k+1} - 2x^k + x^{k-1}}$, которой заменяется итерационное приближение x^{k+1}.

Используем итерационные приближения x^3, x^2 и x^1 для корректировки значения x^3. Итак, $x^{k+1} = x^3$, $x^k = x^2$ и $x^{k-1} = x^1$. Тогда $\overline{x} = \dfrac{x^3x^1 - (x^2)^2}{x^3 - 2x^2 + x^1} \approx 1.34269$ и погрешность $\overline{x}$, равная $|\overline{x} - x^*| \approx 0.018$, меньше погрешности итерационного приближения x^3, которая равна $\left|x^3 - x^*\right| \approx 0.067$.

3. Метод Ньютона.

Используем метод Ньютона для нахождения корня x^* уравнения $x^3 - x - 1 = 0$. Формула для вычисления элементов последовательности итерационных приближений имеет вид

$$x^{k+1} = x^k - \frac{(x^k)^3 - x^k - 1}{3(x^k)^2 - 1},\ k = 0, 1, \ldots .$$

Результаты расчётов приведены в таблице

k	x^k	$\lvert x^k - x^*\rvert$
0	1.1	0.23
1	1.39239544	0.07
2	1.32862605	0.004
3	1.32473211	0.00001
4	1.32471796	0.000000002
5	1.324717957	

Изменения значений погрешности итерационных приближений соответствуют квадратичной скорости сходимости итерационных приближений к корню x^* в методе Ньютона.

3.7.2 Решение системы нелинейных уравнений

Рассмотрим систему двух нелинейных уравнений вида

$$\begin{cases} F(x,\, y) &= x^2 + y^2 - 4 = 0; \\ G(x,\, y) &= xy - 1 = 0. \end{cases} \tag{3.7}$$

На Рис.3.9 изображены линии $x^2 + y^2 - 4 = 0$ и $xy - 1 = 0$. Система нелинейных уравнений (3.7) имеет 4 корня.

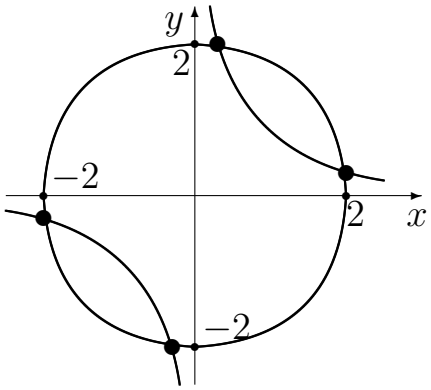

*Рис.*3.9

Используя метод Ньютона, вычислим один из этих корней. Выберем в качестве начального приближения $x^0 = 2$ и $y^0 = 0$.

Линеаризованные уравнения (3.7) в общей форме записи имеют вид

$$\begin{cases} \dfrac{\partial F(x^k,\, y^k)}{\partial x}\Delta x^k + \dfrac{\partial F(x^k,\, y^k)}{\partial y}\Delta y^k &= -F(x^k,\, y^k), \\ \dfrac{\partial G(x^k,\, y^k)}{\partial x}\Delta x^k + \dfrac{\partial G(x^k,\, y^k)}{\partial y}\Delta y^k &= -G(x^k,\, y^k), \end{cases} \qquad k = 0,\, 1,\, \ldots\,.$$

Частные производные функций $F(x,\, y)$ и $G(x,\, y)$ равны, соответственно, $\dfrac{\partial F}{\partial x} = 2x$, $\dfrac{\partial F}{\partial y} = 2y$, $\dfrac{\partial G}{\partial x} = y$, $\dfrac{\partial G}{\partial y} = x$.

Следовательно, на каждой итерации для нахождения приращений Δx^k и Δy^k нужно решать систему двух линейных уравнений

$$\begin{cases} 2x^k\Delta x^k + 2y^k\Delta y^k &= 4 - (x^k)^2 - (y^k)^2, \\ y^k\Delta x^k + x^k\Delta y^k &= 1 - x^k y^k. \end{cases}$$

Следующее итерационное приближение вычисляется по формулам $x^{k+1} = x^k + \Delta x^k$ и $y^{k+1} = y^k + \Delta y^k$.

Результаты расчётов приведены в таблице

k	x^k	y^k	$F(x^k, y^k)$	$G(x^k, y^k)$	Δx^k	Δy^k
0	2	0	0	-1	0	0.5
1	2	0.5	0.25	0	-0.67	0.017
2	1.93	0.517	-0.0077	-0.0022	...	...

Значения функций $F(x^k, y^k)$ и $G(x^k, y^k)$ свидетельствуют о быстрой сходимости итерационного процесса.

Глава 4

Интерполирование и приближение функций

Пусть на отрезке $[a;\ b]$ задан набор точек $x_k,\ k = \overline{0,\ n}$, которые называются узлами интерполирования. В этих точках $a = x_0 < x_1 < \ldots < x_n = b$ пусть заданы значения некоторой функции $f(x_k) = f_k,\ k = \overline{0,\ n}$. Задача состоит в том, чтобы построить такую легко вычисляемую функцию $\Phi(x)$, которая приближает с заданной точностью значения функции $f(x)$ для любого $x \in [a;\ b]$. Функцию $\Phi(x)$ называют интерполянтой.

Интерполянту $\Phi(x)$ представим в виде линейной комбинации базисных функций $\varphi_l(x),\ l = \overline{0,\ m}$, то есть

$$\Phi(x) = \sum_{l=0}^{m} a_l \varphi_l(x).$$

В качестве базисных функций можно использовать, например, степенные функции $\varphi_l(x) = x^l$.

4.1 Интерполирование алгебраическими многочленами

Пусть число используемых базисных функций $\varphi_l(x) = x^l$ равно числу заданных значений функции $f(x)$, то есть выполнено равенство $n = m$, означающее что интерполянта $\Phi(x)$ является полиномом степени n. Потребуем, чтобы в узлах интерполирования, то есть в точках x_k, значения интерполянты $\Phi(x_k)$ совпадали со значениями функции $f(x_k)$. Тогда, для определения коэффициентов a_l, являющихся параметрами интерполянты, получаем сле-

дующую систему линейных алгебраических уравнений

$$\sum_{l=0}^{n} a_l \varphi_l(x_k) = f(x_k),\ k = \overline{0,n}. \tag{4.1}$$

Решение системы (4.1) существует и единственно, если определитель матрицы, составленной из коэффициентов $\varphi_l(x_k)$, отличен от нуля. То есть,

$$\begin{vmatrix} \varphi_0(x_0) & \varphi_1(x_0) & \dots & \varphi_n(x_0) \\ \varphi_0(x_1) & \varphi_1(x_1) & \dots & \varphi_n(x_1) \\ & \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot & & \\ \varphi_0(x_n) & \varphi_1(x_n) & \dots & \varphi_n(x_n) \end{vmatrix} \neq 0. \tag{4.2}$$

Так как $\varphi_l(x) = x^l$, то (4.2) совпадает с определителем Вандермонда

$$\begin{vmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ & \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot & & \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{vmatrix} = \prod_{0 \leqslant k < l \leqslant n} (x_l - x_k) \neq 0.$$

Отсюда следует, что все узлы интерполирования должны быть различными. Тогда решение системы (4.1) существует и единственно.

Используются различные формы записи интерполянты $\Phi(x)$. Представление, в котором выделены значения $f_k = f(x_k)$, называют интерполяционным многочленом Лагранжа:

$$\Phi(x) = L_n(x) = \sum_{k=0}^{n} C_k(x) f_k,$$

где

$$C_k(x) = \frac{\prod\limits_{\substack{i=0 \\ i \neq k}}^{n} (x - x_i)}{\prod\limits_{\substack{i=0 \\ i \neq k}}^{n} (x_k - x_i)}.$$

Интерполирование с кратными узлами

Пусть в узлах интерполирования x_k известны не только значения функции $f_k = f(x_k)$, но и значения всех её производных $f^{(i)}(x_k)$ до $(N_k - 1)$ порядка $(k = \overline{0,\ n}, \quad i = \overline{0,\ N_k - 1})$. Общее число заданных данных равно $\sum\limits_{k=0}^{n} N_k$.

В полиноме $H_m(x) = a_0 + a_1 x + \ldots + a_m x^m$ число коэффициентов $a_0, \ldots, a_m$ равно $(m+1)$. Выберем параметры $a_0, a_1, \ldots, a_m$ так, чтобы выполнялись равенства:

$$H_m^{(i)}(x_k) = f^{(i)}(x_k), \quad при \quad k = \overline{0,\, n}, \quad i = \overline{0,\, N_k - 1}, \tag{4.3}$$

где $m = N_0 + N_1 + \ldots + N_n - 1$.

Многочлен, полученный в результате решения системы линейных алгебраических уравнений (4.3), называется интерполяционным полиномом Эрмита.

Покажем, что система (4.3) разрешима единственным образом. Рассмотрим однородную систему уравнений:

$$H_m^{(i)}(x_k) = 0, \quad при \quad k = \overline{0,\, n},\ i = \overline{0,\, N_k - 1}, \tag{4.4}$$

где $m = N_0 + N_1 + \ldots + N_n - 1$.

Из уравнений (4.4) следует, что x_k являются корнями кратности N_k полинома Эрмита $H_m(x)$. Общее число корней с учетом их кратности равно $N_0 + N_1 + \ldots + N_n$, что на единицу больше степени $m = N_0 + N_1 + \ldots + N_n - 1$ полинома $H_m(x)$. Это возможно только в случае, когда полином тождественно равен нулю, то есть $a_0 = a_1 = \ldots = a_m = 0$ — равны нулю все его коэффициенты. Итак, однородная система (4.4) имеет только тривиальное решение. Следовательно, решение неоднородной системы (4.3) существует и единственно.

Формула, определяющая интерполяционный полином Эрмита, имеет вид (см. [6]):

$$H_m(x) = \sum_{k=0}^{n} \sum_{i=0}^{N_k-1} \sum_{l=0}^{N_k-1-i} \frac{f^{(i)}(x_k)}{i!l!} \left((x - x_k)^{i+l} \prod_{\substack{j=0 \\ j\neq k}}^{n} (x - x_j)^{N_j} \right) \cdot \left(\frac{d^l}{dx^l} \prod_{\substack{j=0 \\ j\neq k}}^{n} (x - x_j)^{-N_j} \right)_{x=x_k}.$$

Если все параметры $N_k = 1$, то две внутренние суммы дают одно слагаемое с $i = l = 0$ и $H_m(x)$ переходит в интерполяционный многочлен Лагранжа.

Если все $N_k = 2$, то $H_m(x)$ принимает вид

$$H_m(x) = \sum_{k=0}^{n} \left(\left((x - x_k) f^{(1)}(x_k) + \left(1 - 2 \sum_{\substack{j=0 \\ j\neq k}}^{n} \frac{x - x_k}{x_k - x_j} \right) f_k \right) \cdot \right.$$

$$\cdot \prod_{\substack{j=0 \\ j \neq k}}^{n} \left(\frac{x - x_j}{x_k - x_j} \right)^2 \Bigg).$$

Сходимость интерполяционного процесса

Введем на отрезке $[a;\, b]$ последовательность наборов узлов интерполирования $\Omega_n = \{x_k^{(n)},\, Ck = \overline{0,n}\} : \Omega_1 = \{x_0^1,\, x_1^1\},\ \ldots,\ \Omega_n = \{x_0^n,\, \ldots,\, x_n^n\},\ \ldots$. Пусть для каждого Ω_n имеется соответствующий набор значений приближаемой функции $f(x_k) = f_k,\ k = \overline{0,n}$. Для каждого из наборов узлов интерполирования Ω_n построим интерполяционный многочлен Лагранжа $L_n(x)$.

Введём определения.

Определение. Интерполяционный многочлен Лагранжа сходится в точке x к функции $f(x)$, если существует предел $\lim\limits_{n\to\infty} L_n(x) = f(x)$.

Определение. Интерполяционный многочлен Лагранжа сходится равномерно к функции $f(x)$ на отрезке $[a;\, b]$, если $\max\limits_{x\in[a;\,b]} |L_n(x) - f(x)| \overset{n\to\infty}{\longrightarrow} 0$.

Пример 4.1. Рассмотрим функцию $f(x) = |x|$ на отрезке $[-1;\, 1]$. Введём на этом отрезке набор узлов интерполирования $\Omega_n = \{x_k = \pm\frac{k}{n},\ k = \overline{0,\, n}\}$. Вычислим $f(x_k)$ и построим интерполяционный многочлен Лагранжа $L_{2n}(x)$.

Известно [6], что для любой точки $x \in (-1;\, 1)$ кроме $x = 0$ нет сходимости интерполяционного многочлена Лагранжа к функции $f(x)$, то есть

$$\lim_{n\to\infty} |L_{2n}(x) - f(x)| \nrightarrow 0.$$

Верно, также, следующее утверждение:

Теорема 4.1. Для любой непрерывной на отрезке $[a;\, b]$ функции $f(x)$ существует такая последовательность узлов интерполирования Ω_n, что соответствующая последовательность многочленов Лагранжа сходится равномерно к функции $f(x)$ на отрезке $[a;\, b]$.

Из-за такой неоднозначности поведения интерполяционных многочленов Лагранжа, как правило, не используют интерполирование многочленами высокой степени. При необходимости интерполирования на протяженном отрезке $[a;\, b]$ его делят на частичные сегменты и на каждом сегменте строят интерполяционные многочлены не высокой степени. Такую процедуру называют кусочно-полиномиальным интерполированием.

4.2 Интерполирование сплайнами

Разобьем отрезок $[a;\ b]$ точками $a = x_0 < x_1 < \ldots < x_n = b$ на n сегментов $[x_{i-1};\ x_i],\ i = \overline{1,\ n}$.

Пусть, в узлах интерполирования x_i заданы значения некоторой функции $f(x_i) = f_i,\ i = \overline{0,\ n}$.

Рассмотрим функцию $S_m(x)$, которая на каждом сегменте $[x_{i-1};\ x_i]$ является полиномом степени m :

$$S_m(x) = P_{im}(x) = a_{i0} + a_{i1}x + \ldots + a_{im}x^m, \quad x \in [x_{i-1};\ x_i],\ i = \overline{1,\ n}.$$

Потребуем, чтобы в точках $x_i (i = \overline{1,\ n})$ производные до $(m-1)$ порядка функции $S_m(x)$ были непрерывны:

$$P^{(l)}_{im}(x_i) = P^{(l)}_{(i+1)m}(x_i), \quad l = \overline{0,\ m-1},\ i = \overline{1,\ n-1},$$

и значения функции $S_m(x)$ в точках $x_i (i = \overline{0,\ n})$ были равны значениям функции $f(x)$:

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

Функцию $S_m(x)$, удовлетворяющую всем этим условиям, называют интерполяционным сплайном степени m на отрезке $[a;\ b]$.

Функция $S_m(x)$ определяется заданием коэффициентов $a_{ij} (i = \overline{1,\ n},\ j = \overline{0,\ m})$, число которых равно $nm + n$. Количество условий в точках x_i на производные и значения интерполяционного сплайна степени m равно $nm + n - (m-1)$. Дополнительные $(m-1)$ условия, обычно, задают на концах отрезка $[a;\ b]$.

В итоге, для определения коэффициентов a_{ij}, получают систему из $n(m+1)$ линейных алгебраических уравнений.

4.2.1 Интерполирование кубическими сплайнами

Рассмотрим функцию $S_3(x)$, которую называют кубическим сплайном. Запишем $S_3(x)$ в виде:

$$S_3(x) = P_{i3}(x) = a_i + b_i(x - x_i) + \frac{c_i}{2}(x - x_i)^2 + \frac{d_i}{6}(x - x_i)^3,$$
$$x \in [x_{i-1};\ x_i],\ i = \overline{1,\ n}. \quad (4.5)$$

Коэффициенты $a_i,\ b_i,\ c_i,\ d_i,\ i = \overline{1,\ n}$ являются параметрами кубического сплайна $S_3(x)$. Количество параметров равно $4n$. Задание числовых значений этих коэффициентов определяет кубический сплайн.

Следствием условия совпадения значений функции $f(x)$ и кубического сплайна $S_3(x)$ в узлах интерполирования $x_i\,(i = \overline{1,\,n})$, являются равенства $a_i = f_i,\ i = \overline{1,\,n}$.

Первая производная кубического сплайна равна $S_3'(x) = b_i + c_i(x - x_i) + \dfrac{d_i}{2}(x - x_i)^2$. Условие непрерывности $S_3'(x)$ во внутренних узлах интерполирования приводит к равенствам:

$$b_i = b_{i+1} - c_{i+1}h_{i+1} + \frac{d_{i+1}}{2}h_{i+1}^2,\ i = \overline{1,\,n-1},$$
$$\text{где}\quad h_{i+1} = x_{i+1} - x_i,\ i = \overline{0,\,n-1}. \quad (4.6)$$

Вторая производная кубического сплайна равна $S_3''(x) = c_i + d_i(x - x_i)$. Условие непрерывности $S_3''(x)$ во внутренних узлах интерполирования дает систему равенств:

$$c_i = c_{i+1} - d_{i+1}h_{i+1},\ i = \overline{1,\,n-1}. \quad (4.7)$$

Условие непрерывности $S_3(x)$ во внутренних узлах интерполирования приводит к равенствам:

$$f_i = f_{i+1} - b_{i+1}h_{i+1} + \frac{c_{i+1}}{2}h_{i+1}^2 - \frac{d_{i+1}}{6}h_{i+1}^3,\ i = \overline{1,\,n-1}. \quad (4.8)$$

Итак, всего имеем $(4n-2)$ уравнений для определения $4n$ коэффициентов $a_i,\ b_i\,c_i\,d_i,\ i = \overline{1,\,n}$.

В качестве двух дополнительных условий используем требование нулевой кривизны кубического сплайна на концах отрезка интерполирования: $S_3''(x_0) = S_3''(x_n) = 0$. Используя выражение для $S_3''(x) = c_i + d_i(x - x_i)$, получим, что условие $S_3''(x_n) = 0$ приводит к равенству $c_n = 0$, а из условия $S_3''(x_0) = 0$ следует равенство $c_1 - d_1h_1 = 0$, которое совпадает с (4.7) при $i = 0$, если ввести дополнительный коэффициент $c_0 = 0$.

Исключим из равенств (4.8) коэффициенты b_i и d_i. Для этого приведём (4.8) к виду:

$$b_{i+1} - \frac{c_{i+1}}{2}h_{i+1} + \frac{d_{i+1}}{6}h_{i+1}^2 = \frac{f_{i+1} - f_i}{h_{i+1}},\ i = \overline{1,\,n-1}. \quad (4.9)$$

Заменим в (4.9) индекс i на $i - 1$:

$$b_i - \frac{c_i}{2}h_i + \frac{d_i}{6}h_i^2 = \frac{f_i - f_{i-1}}{h_i},\ i = \overline{1,\,n-1}. \quad (4.10)$$

Вычтем (4.10) из (4.9). Заменим в разности комбинацию $b_{i+1} - b_i$, используя (4.6), и все коэффициенты d_i, используя (4.7). В результате получим следую-

щую систему линейных алгебраических уравнений для вычисления коэффициентов c_i :

$$\begin{cases} h_i c_{i-1} + 2(h_i + h_{i+1})c_i + h_{i+1}c_{i+1} = 6\left(\dfrac{f_{i+1} - f_i}{h_{i+1}} - \dfrac{f_i - f_{i-1}}{h_i}\right), \; i = \overline{1,\, n-1}; \\ c_0 = c_n = 0. \end{cases} \tag{4.11}$$

Решив методом прогонки систему уравнений (4.11), коэффициенты b_i и d_i вычисляют по формулам:

$$\begin{cases} d_i &= \dfrac{c_i - c_{i-1}}{h_i}; \\ b_i &= \dfrac{c_i h_i}{2} - \dfrac{d_i h_i^2}{6} + \dfrac{f_i - f_{i-1}}{h_i}, \quad i = \overline{1,\, n}. \end{cases}$$

Отметим, что при изменении одного значения f_i приближаемой функции меняются все коэффициенты кубического сплайна.

4.2.2 Сходимость процесса интерполирования кубическими сплайнами

Покажем [3], что интерполирование кубическими сплайнами $S_3(x)$ является сходящимся процессом, то есть с увеличением числа узлов интерполирования, соответствующая последовательность сплайнов $S_3(x)$ сходится к функции $f(x)$.

Пусть узлами интерполирования являются точки $\Omega_n = \{x_i = a + ih,\; i = \overline{0,\, n},\; h = \dfrac{b-a}{n}\}$, равномерно распределённые на отрезке $[a;\, b]$.

Введем обозначения $||g(x)||_{C[a;\,b]} = \max\limits_{x \in [a;\,b]} |g(x)|$, $||g_i||_{C(\Omega_n)} = \max\limits_i |g_i|$

и $M = ||f^{(4)}(x)||_{C[a;\,b]}$.

Докажем вспомогательную лемму.

Лемма. Пусть функция $f(x) \in C^4[a;\, b]$ и $f''(a) = f''(b) = 0$. Пусть $S_3(x)$ соответствующий функции $f(x)$ кубический сплайн, построенный на множестве Ω_n. Тогда верно неравенство

$$||f''(x_i) - S_3''(x_i)||_{C(\Omega_n)} \leqslant \frac{3}{4} M h^2. \tag{4.12}$$

Доказательство. Система уравнений (4.11) на множестве Ω_n принимает вид:

$$\begin{cases} c_{i-1} + 4c_i + c_{i+1} = 6 f_{\bar{x}x,i}, \quad i = \overline{1,\, n-1}; \\ c_0 = c_n = 0, \end{cases} \tag{4.13}$$

где $f_{\overline{x}x,i} = \frac{f_{i+1}-2f_i+f_{i-1}}{h^2}$ — вторая разностная производная функции $f(x)$ в точке x_i.

Введём обозначение $z_i = S_3''(x_i) - f''(x_i)$. Из соотношения $S_3''(x) = c_i + d_i(x-x_i)$ следует, что $S_3''(x_i) = c_i$. Тогда c_i можно записать в виде $c_i = z_i + f_i''$. Подставив это выражение для c_i в (4.13), получим:

$$\begin{cases} z_{i-1} + 4z_i + z_{i+1} = \psi_i, \quad i = \overline{1,\, n-1}; \\ z_0 = z_n = 0, \end{cases}$$

где $\psi_i = 6f_{\overline{x}x,i} - (f_{i-1}'' + 4f_i'' + f_{i-1}'')$.

Запишем уравнение для z_i в виде

$$4z_i = -z_{i-1} - z_{i+1} + \psi_i, \quad i = \overline{1,\, n-1}.$$

Взяв значения z_i по модулю, получим:

$$4|z_i| \leqslant |z_{i-1}| + |z_{i+1}| + |\psi_i| \leqslant 2\max_{i=\overline{0,\,n}} |z_i| + \max_{i=\overline{1,\,n-1}} |\psi_i| = \\ = 2||z_i||_{C(\Omega_n)} + ||\psi_i||_{C(\Omega_n)}, \quad i = \overline{1,\, n-1}.$$

Выбрав в левой части неравенства максимальное значение $|z_i|$, приходим к неравенству

$$4||z_i||_{C(\Omega_n)} \leqslant 2||z_i||_{C(\Omega_n)} + ||\psi_i||_{C(\Omega_n)}.$$

Отсюда следует, что

$$||z_i||_{C(\Omega_n)} \leqslant \frac{1}{2}||\psi_i||_{C(\Omega_n)}. \tag{4.14}$$

Теперь получим оценку для $||\psi_i||_{C(\Omega_n)}$.

Запишем ψ_i в виде

$$\psi_i = 6f_{\overline{x}x,i} - (f_{i-1}'' + 4f_i'' + f_{i-1}'') = 6(f_{\overline{x}x,i} - f_i'') - \frac{f_{i-1}'' - 2f_i'' + f_{i+1}''}{h^2}h^2 = \\ = 6(f_{\overline{x}x,i} - f_i'') - f_{\overline{x}x,i}''h^2.$$

Вторая разностная производная $f_{\overline{x}x,i}$ равна

$$f_{\overline{x}x,i} = \frac{1}{h^2}(f_{i-1} - 2f_i + f_{i+1}) = \left\{ f_{i\pm1} = f(x_i \pm h) = f_i \pm f_i'h + f_i''\frac{h^2}{2} \pm \right. \\ \left. \pm f_i^{(3)}\frac{h^3}{6} + f^{(4)}(\xi_i^{\pm})\frac{h^4}{24}, \quad \xi_i^+,\, \xi_i^- \in [x_{i-1};\, x_{i+1}] \right\} = \\ = f_i'' + f^{(4)}(\xi_i^-)\frac{h^2}{24} + f^{(4)}(\xi_i^+)\frac{h^2}{24}.$$

Аналогично получаем, что $f''_{\overline{x}x,i}$ представима в виде

$$f''_{\overline{x}x,i} = \frac{1}{h^2}(f''_{i-1} - 2f''_i + f''_{i+1}) = \Big\{ f''_{i\pm1} = f''(x_i \pm h) = f''_i \pm f^{(3)}_i h + \\ + f^{(4)}(\zeta_i^{\pm})\frac{h^2}{2}, \quad \zeta_i^+,\, \zeta_i^- \in [x_{i-1};\, x_{i+1}] \Big\} = \frac{1}{2}f^{(4)}(\zeta_i^-) + \frac{1}{2}f^{(4)}(\zeta_i^+).$$

Подставив полученные выражения для $f_{\overline{x}x,i}$ и $f''_{\overline{x}x,i}$ в ψ_i, получим

$$\psi_i = f^{(4)}(\xi_i^-)\frac{h^2}{4} + f^{(4)}(\xi_i^+)\frac{h^2}{4} - f^{(4)}(\zeta_i^-)\frac{h^2}{2} - f^{(4)}(\zeta_i^+)\frac{h^2}{2}, \quad \xi_i^{\pm},\, \zeta_i^{\pm} \in [x_{i-1};\, x_{i+1}].$$

Отсюда следует, что

$$|\psi_i| \leqslant |f^{(4)}(\xi_i^-)|\frac{h^2}{4} + |f^{(4)}(\xi_i^+)|\frac{h^2}{4} + |f^{(4)}(\zeta_i^-)|\frac{h^2}{2} + |f^{(4)}(\zeta_i^+)|\frac{h^2}{2} \leqslant \\ \leqslant \frac{3}{2}\max_{x\in[a;\,b]}|f^{(4)}(x)|h^2 = \frac{3}{2}Mh^2, \quad i = \overline{1,\, n-1}.$$

Выбрав в левой части максимум $|\psi_i|$, имеем неравенство

$$||\psi_i||_{C(\Omega_n)} \leqslant \frac{3}{2}Mh^2.$$

Подставляя это неравенство в (4.14), получаем (4.12). □

Докажем теорему о сходимости процесса интерполирования кубическими сплайнами.

Теорема 4.2. Пусть $f(x) \in C^4[a;\, b]$ и $f''(a) = f''(b) = 0$. Пусть $S_3(x)$ соответствующий функции $f(x)$ кубический сплайн, построенный на множестве Ω_n. Тогда верны неравенства:

$$||f(x) - S_3(x)||_{C[a;\,b]} < Mh^4; \tag{4.15}$$

$$||f'(x) - S'_3(x)||_{C[a;\,b]} \leqslant Mh^3; \tag{4.16}$$

$$||f''(x) - S''_3(x)||_{C[a;\,b]} \leqslant Mh^2. \tag{4.17}$$

Доказательство. **(1).** Покажем, что верно неравенство (4.17). Из (4.5) и (4.7) получаем, что:

$$S_3''(x) = c_i + d_i(x - x_i) = c_i + \frac{c_i - c_{i-1}}{h}(x - x_i) = c_i\frac{x - (x_i - h)}{h} - \\ - c_{i-1}\frac{x - x_i}{h} = c_i\frac{x - x_{i-1}}{h} + c_{i-1}\frac{x_i - x}{h}, \quad x \in [x_{i-1};\, x_i], \quad i = \overline{1,\, n}.$$

Тогда,

$$f''(x) - S_3''(x) = f''(x) - c_i\frac{x - x_{i-1}}{h} - c_{i-1}\frac{x_i - x}{h} = f''(x)\frac{x - x_{i-1}}{h} + \\ + f''(x)\frac{x_i - x}{h} - f_i''\frac{x - x_{i-1}}{h} + (f_i'' - c_i)\frac{x - x_{i-1}}{h} - f_{i-1}''\frac{x_i - x}{h} + \\ + (f_{i-1}'' - c_{i-1})\frac{x_i - x}{h} = (f''(x) - f_i'')\frac{x - x_{i-1}}{h} + (f''(x) - f_{i-1}'')\frac{x_i - x}{h} + \\ + (f_i'' - c_i)\frac{x - x_{i-1}}{h} + (f_{i-1}'' - c_{i-1})\frac{x_i - x}{h}.$$

Отсюда следует, что

$$|f''(x) - S_3''(x)| \leqslant \left|(f''(x) - f_i'')\frac{x - x_{i-1}}{h} + (f''(x) - f_{i-1}'')\frac{x_i - x}{h}\right| + \\ + \frac{x - x_{i-1}}{h}\,|f_i'' - c_i| + \frac{x_i - x}{h}\,|f_{i-1}'' - c_{i-1}|\,. \quad (4.18)$$

Для двух слагаемых из (4.18), с учётом (4.12), верна оценка

$$\frac{x - x_{i-1}}{h}|f_i'' - c_i| + \frac{x_i - x}{h}|f_{i-1}'' - c_{i-1}| \leqslant \left(\frac{x - x_{i-1}}{h} + \frac{x_i - x}{h}\right) \cdot \\ \cdot \max_i |f_i'' - c_i| = ||f''(x_i) - S_3''(x_i)||_{C(\Omega_n)} \leqslant \frac{3}{4}Mh^2. \quad (4.19)$$

Представим разность $f''(x) - f''(x_i)$ в следующем виде

$$f''(x) - f''(x_i) = f''(x) - \left(f''(x) + f^{(3)}(x)(x_i - x) + f^{(4)}(\xi_i)\frac{(x_i - x)^2}{2}\right) = \\ = -f^{(3)}(x)(x_i - x) - f^{(4)}(\xi_i)\frac{(x_i - x)^2}{2}, \quad \xi_i \in [x_{i-1};\, x_i].$$

Для $f''(x) - f''(x_{i-1})$ верно аналогичное представление

$$f''(x) - f''(x_{i-1}) = f''(x) - \Big(f''(x) + f^{(3)}(x)(x_{i-1} - x) + f^{(4)}(\zeta_i)\cdot \\ \cdot\frac{(x_{i-1} - x)^2}{2}\Big) = -f^{(3)}(x)(x_{i-1} - x) - f^{(4)}(\zeta_i)\frac{(x_{i-1} - x)^2}{2}, \quad \zeta_i \in [x_{i-1};\, x_i].$$

Тогда:

$$\left|(f''(x)-f_i'')\frac{x-x_{i-1}}{h}+(f''(x)-f_{i-1}'')\frac{x_i-x}{h}\right|=\left|-f^{(3)}(x)\frac{(x-x_{i-1})}{h}\cdot\right.$$
$$\cdot(x_i-x)-f^{(3)}(x)\frac{(x_{i-1}-x)(x_i-x)}{h}-f^{(4)}(\xi_i)\frac{(x-x_{i-1})(x_i-x)^2}{2h}-$$
$$\left.-f^{(4)}(\zeta_i)\frac{(x_i-x)(x_{i-1}-x)^2}{2h}\right|=\left|f^{(4)}(\xi_i)\frac{(x-x_{i-1})(x_i-x)^2}{2h}+\right.$$
$$\left.+f^{(4)}(\zeta_i)\frac{(x_i-x)(x_{i-1}-x)^2}{2h}\right|=(x-x_{i-1})(x_i-x)\left|f^{(4)}(\xi_i)\frac{x_i-x}{2h}+\right.$$
$$\left.+f^{(4)}(\zeta_i)\frac{x-x_{i-1}}{2h}\right|\leqslant\max((x-x_{i-1})(x_i-x))\cdot\max\left(\left|f^{(4)}(\xi_i)\frac{x_i-x}{2h}\right|+\right.$$
$$\left.+\left|f^{(4)}(\zeta_i)\frac{x-x_{i-1}}{2h}\right|\right)\leqslant\left(\frac{x_{i-1}+x_i}{2}-x_{i-1}\right)\left(x_i-\frac{x_{i-1}+x_i}{2}\right)M=\frac{Mh^2}{4}. \tag{4.20}$$

Подставляя (4.20) и (4.19) в (4.18), имеем

$$|f''(x)-S_3''(x)|\leqslant Mh^2,\quad\forall\,x\in[x_{i-1};\,x_i],\quad i=\overline{1,\,(n-1)}.$$

Выбирая x, при котором достигается максимум, получаем неравенство (4.17)

$$||f''(x)-S_3''(x)||_{C[a;\,b]}\leqslant Mh^2.$$

(2). Докажем неравенство (4.16). Рассмотрим разность $f(x)-S_3(x)$. Так как $f(x_i)-S_3(x_i)=0,\quad i=\overline{0,\,n}$, то существуют такие точки $\xi_i\in[x_{i-1};\,x_i],\quad i=\overline{1,\,n}$, что $f'(\xi_i)-S_3'(\xi_i)=0$ (теорема Ролля).

Тогда, используя формулу Лагранжа, получаем

$$f'(x)-S_3'(x)=(f'(x)-S_3'(x))-(f'(\xi_i)-S_3'(\xi_i))=(f''(\zeta_i)-S_3''(\zeta_i))\cdot(x-\xi_i),$$
$$x,\,\xi_i,\,\zeta_i\in[x_{i-1};\,x_i],\quad i=\overline{1,\,n}.$$

Отсюда, с учётом (4.17), следует, что

$$|f'(x)-S_3'(x)|\leqslant|f''(\zeta_i)-S_3''(\zeta_i)|\cdot h\leqslant Mh^3.$$

Взяв в левой части $x\in[a;\,b]$, в котором достигается максимум, получим невенство (4.16)

$$||f'(x)-S'(x)||_{C[a;\,b]}\leqslant Mh^3.$$

(3). Рассмотрим на отрезке $[x_{i-1};\,x_i],\quad i=\overline{1,\,n}$ вспомогательную функцию

$$g(t)=f(t)-S_3(t)-K(t-x_{i-1})(t-x_i),\quad t\in[x_{i-1};\,x_i],$$

где K — константа.

Выберем произвольное $x \in (x_{i-1};\, x_i)$. Подберём K так, чтобы $g(x) = 0$, то есть

$$f(x) - S_3(x) - K(x - x_{i-1})(x - x_i) = 0.$$

Отсюда получаем, что $K = \dfrac{f(x) - S_3(x)}{(x - x_{i-1})(x - x_i)}$.

Итак,

$$g(t) = f(t) - S_3(t) - \frac{f(x) - S_3(x)}{(x - x_{i-1})(x - x_i)}(t - x_{i-1})(t - x_i).$$

Функция $g(t)$ дважды дифференцируема на $(x_{i-1};\, x_i)$ и $g(x_{i-1}) = g(x_i) = g(x) = 0$ — обращается в нуль в трех точках. Следовательно (теорема Ролля), непрерывная первая производная равна нулю в двух точках и существует такая точка $\xi_i \in (x_{i-1};\, x_i)$, что $g''(\xi_i) = 0$.

Так как $g''(t) = f''(t) - S_3''(t) - 2K$, то

$$g''(\xi_i) = f''(\xi_i) - S_3''(\xi_i) - 2\frac{f(x) - S_3(x)}{(x - x_{i-1})(x - x_i)} = 0.$$

Отсюда следует, что

$$f(x) - S_3(x) = \frac{f''(\xi_i) - S_3''(\xi_i)}{2}(x - x_{i-1})(x - x_i).$$

Переходя к модулю, получим

$$|f(x) - S_3(x)| = \frac{1}{2}|f''(\xi_i) - S_3''(\xi_i)| \cdot |(x - x_{i-1})(x - x_i)| \leqslant \frac{1}{2} \cdot Mh^2 \cdot \frac{h^2}{4}.$$

В силу произвольности выбора $i = \overline{1,\, n}$ и x, неравенство верно для $x \in [a;\, b]$, в котором достигается максимум левой части. Следовательно,

$$||f(x) - S_3(x)||_{C[a;\, b]} \leqslant \frac{Mh^4}{8}$$

и верно неравенство (4.15). Теорема доказана. □

4.3 Наилучшее приближение в гильбертовом пространстве

Пусть $\mathbb{H}$ — линейное нормированное пространство функций, а φ_i $(i = \overline{0,\, n})$ — линейно независимые элементы $\mathbb{H}$. Заданному элементу $f \in \mathbb{H}$ сопоставим линейную комбинацию

$$\varphi = c_0\varphi_0 + \ldots + c_n\varphi_n, \quad c_i = const, \quad i = \overline{0,\, n}.$$

Определение. Элемент φ, доставляющий минимум норме $\|f-\varphi\|_{\mathbb{H}}$, называется элементом наилучшего приближения.

Покажем на примере пространства $L_2[a;\,b]$, что элемент наилучшего приближения существует и единственен. Определим скалярное произведение функций f и g в L_2 как

$$(g,\,f)_{L_2}=\int\limits_a^b f(x)g(x)\,dx \quad \text{и, соответственно,} \qquad \|f\|_{L_2}=\left(\int\limits_a^b f^2(x)\,dx\right)^{\frac{1}{2}}.$$

Преобразуем выражение для $\|f-\varphi\|_{L_2}$:

$$\|f-\varphi\|_{L_2}^2=(f-\varphi,\,f-\varphi)_{L_2}=\left(f-\sum_{l=0}^{n}c_l\varphi_l,\,f-\sum_{k=0}^{n}c_k\varphi_k\right)_{L_2}=$$

$$=(f,\,f)_{L_2}-\sum_{l=0}^{n}c_l\,(\varphi_l,\,f)_{L_2}-\sum_{k=0}^{n}c_k\,(\varphi_k,\,f)_{L_2}+\sum_{l=0}^{n}\sum_{k=0}^{n}c_lc_k\,(\varphi_l,\,\varphi_k)_{L_2}=$$

$$=\|f\|_{L_2}^2-2\sum_{l=0}^{n}c_l\,(\varphi_l,\,f)_{L_2}+\sum_{l=0}^{n}\sum_{k=0}^{n}c_lc_k\,(\varphi_l,\,\varphi_k)_{L_2}.$$

Введем обозначения $f_l=(\varphi_l,\,f)_{L_2}=\int\limits_a^b f(x)\varphi_l(x)\,dx$ и $a_{lk}=(\varphi_l,\,\varphi_k)_{L_2}=\int\limits_a^b\varphi_l(x)\varphi_k(x)\,dx$. Пусть $c_l,\ f_l\,(l=\overline{0,\,n})$ являются компонентами векторов $\overline{c}=(c_0,\,c_1,\ldots,\,c_n)^T$, $\overline{f}=(f_0,\,f_1,\ldots,\,f_n)^T$ и $a_{lk}\,(l,k=\overline{0,\,n})$ являются элементами матрицы A.

Тогда

$$\|f-\varphi\|_{L_2}^2=\|f\|_{L_2}^2-2\sum_{l=0}^{n}c_lf_l+\sum_{l=0}^{n}\sum_{k=0}^{n}c_lc_ka_{lk}=\|f\|_{L_2}^2-2\left(\overline{f},\,\overline{c}\right)+$$

$$+(A\overline{c},\,\overline{c})=\|f\|_{L_2}^2+J\left(\overline{c}\right), \quad (4.21)$$

где $J\left(\overline{c}\right)=(A\overline{c},\,\overline{c})-2\left(\overline{f},\,\overline{c}\right)$.

Таким образом, минимизация $\|f-\varphi\|_{L_2}$ сводится к поиску минимума $J(\overline{c})$ — функции многих переменных.

Из определения элементов a_{lk} следует, что матрицы A симметрична, то есть $a_{lk}=a_{kl}$. Покажем, что матрицы A положительно определена, то есть $(A\overline{c},\,\overline{c})>0, \quad \forall\overline{c}\neq 0$.

Если взять $f \equiv 0$, то равенство (4.21) примет вид $0 \le \|\varphi\|^2_{L_2} = (A\overline{c}, \overline{c})$. Пусть существует вектор $\overline{c} = (c_0, c_1, \ldots, c_n)^T \neq 0$ такой, что $(A\overline{c}, \overline{c}) = 0$. Тогда $\|\varphi\|^2_{L_2} = 0$ и, следовательно, комбинация $\varphi = c_0\varphi_0 + \ldots + c_n\varphi_n = 0$, в которой все φ_i $(i = \overline{0, n})$ — линейно независимы. Это возможно только в случае $c_0 = c_1 = \ldots = c_n = 0$. Таким образом, матрица $A = A^T > 0$.

Верна следующая теорема.

Теорема 4.3. Пусть матрица $A = A^T > 0$ и $\overline{f}$ — заданный вектор. Тогда у функции $J(\overline{c}) = (A\overline{c}, \overline{c}) - 2\left(\overline{f}, \overline{c}\right)$ минимум существует, единственен и вектор $\overline{c}$ реализует этот минимум тогда и только тогда, когда является решением системы линейных уравнений $A\overline{c} = \overline{f}$.

Доказательство. Утверждение об эквивалентности поиска минимума функции $J(\overline{c})$ и решения системы $A\overline{c} = \overline{f}$ было доказано ранее (см. n.1.6.1). Так как матрица $A > 0$, то решение системы линейных алгебраических уравнений $A\overline{c} = \overline{f}$ существует и единственно. Теорема доказана. □

Алгоритм построения элемента наилучшего приближения

Алгоритм построения элемента наилучшего приближения для функции $f \in L_2[a; b]$ состоит в следующем.

(1). Выбираем набор из $(n+1)$ линейно независимых функций φ_k, $k = \overline{0, n}$ из $L_2[a; b]$.

(2). Вычисляем элементы матрицы $A = (a_{kl})$:

$$a_{kl} = \int_a^b \varphi_k(x)\varphi_l(x)\,dx,\ k, l = \overline{0, n}.$$

(3). Вычисляем компоненты $\overline{f} = (f_0, f_1, \ldots, f_n)^T$:

$$f_k = \int_a^b f(x)\varphi_k(x)\,dx, \quad k = \overline{0, n}.$$

(4). Решаем систему линейных уравнений $A\overline{c} = \overline{f}$.

(5). Строим элемент наилучшего приближения:

$$\varphi = c_0\varphi_0 + c_1\varphi_1 + \ldots + c_n\varphi_n.$$

Погрешность элемента наилучшего приближения

Оценим величину отклонения $\|f-\varphi\|_{L_2}$. Для этого докажем лемму.

Лемма. Пусть φ — элемент наилучшего приближения для f.

Тогда $(f-\varphi,\,\varphi)_{L_2} = (f,\,\varphi)_{L_2} - \|\varphi\|^2_{L_2} = 0$.

Доказательство. Преобразуем скалярное произведение

$$(f-\varphi,\,\varphi)_{L_2} = \left(f - \sum_{k=0}^{n} c_k\varphi_k,\, \sum_{l=0}^{n} c_l\varphi_l\right)_{L_2} = \sum_{l=0}^{n} c_l\,(f,\,\varphi_l)_{L_2} -$$
$$- \sum_{k=0}^{n}\sum_{l=0}^{n} c_k c_l\,(\varphi_k,\,\varphi_l)_{L_2} = \sum_{l=0}^{n} c_l f_l - \sum_{k=0}^{n}\sum_{l=0}^{n} c_k c_l a_{kl} =$$
$$= \left(\overline{f},\,\overline{c}\right) - (A\overline{c},\,\overline{c}) = \left(\overline{f} - A\overline{c},\,\overline{c}\right).$$

Так как $\overline{c}$ является решением системы $A\overline{c} = \overline{f}$, то $(f-\varphi,\,\varphi)_{L_2} = 0$. Лемма доказана. □

Величина отклонения $\|f-\varphi\|^2_{L_2}$ равна:

$$\|f-\varphi\|^2_{L_2} = (f-\varphi,\,f-\varphi)_{L_2} = (f-\varphi,\,f)_{L_2} - (f-\varphi,\,\varphi)_{L_2} =$$
$$= (f-\varphi,\,f)_{L_2} = (f,\,f)_{L_2} - (\varphi,\,f)_{L_2} = (f,\,f)_{L_2} - (\varphi,\,\varphi)_{L_2} = \|f\|^2_{L_2} - \|\varphi\|^2_{L_2}.$$

Замечание. Если элементы φ_k — образуют ортонормированную систему, то есть $(\varphi_k,\,\varphi_l) = \delta_{kl}$, то $A = E$. Тогда $c_k = f_k = (f,\,\varphi_k)_{L_2}$ и наилучшее приближение имеет вид $\varphi = \sum\limits_{k=0}^{n} f_k\varphi_k$.

В этом случае коэффициенты c_k называют коэффициентами Фурье, а элемент φ — многочленом Фурье.

Пример 4.1.

Пусть функция $f(x)$ задана в точках x_0, $x_1 = x_0 + h$, $x_2 = x_0 + 2h$. Построим для этой функции элемент наилучшено приближения. Введём обозначения $F_0 = f(x_0)$, $F_1 = f(x_1)$, $F_2 = f(x_2)$.

Реализуем алгоритм построения элемента наилучшего приближения.

(1). Выберем $\varphi_0(x) = 1$ и $\varphi_1(x) = x - x_1, \quad x \in [x_1 - h;\, x_1 + h]$.

(2). Вычислим $a_{kl} = \int\limits_{x_1-h}^{x_1+h} \varphi_k(x)\varphi_l(x)\,dx, \quad k,\,l = 0,\,1:$

$$a_{00} = 2h;$$
$$a_{10} = a_{01} = \int\limits_{x_1-h}^{x_1+h} (x - x_1)\,dx = \frac{(x-x_1)^2}{2}\Big|_{x_1-h}^{x_1+h} = 0;$$
$$a_{11} = \int\limits_{x_1-h}^{x_1+h} (x - x_1)^2\,dx = \frac{(x-x_1)^3}{3}\Big|_{x_1-h}^{x_1+h} = \frac{2h^3}{3}.$$

Итак, матрица $A = \begin{pmatrix} 2h & 0 \\ 0 & \dfrac{2h^3}{3} \end{pmatrix}$.

(3). Вычислим $f_k = \int\limits_{x_1-h}^{x_1+h} f(x)\varphi_k(x)\,dx, \quad k = 0,\, 1.$ Используя квадратурную формулу Симпсона получим:

$$f_0 = \int\limits_{x_1-h}^{x_1+h} f(x)\,dx = \left\{ \int\limits_a^b G(x)\,dx \approx \frac{b-a}{6}\,(G_0 + 4G_1 + G_2) \right\} \approx$$
$$\approx \frac{h}{3}\,(F_0 + 4F_1 + F_2)\,;$$
$$f_1 = \int\limits_{x_1-h}^{x_1+h} f(x)(x - x_1)\,dx = \left\{ \int\limits_a^b G(x)\,dx \approx \frac{b-a}{6}\,(G_0 + 4G_1 + G_2) \right\} \approx$$
$$\approx \frac{h}{3}\,(F_0 \cdot (-h) + 4F_1 \cdot 0 + F_2 \cdot h) = \frac{h^2}{3}\,(F_2 - F_0)\,.$$

(4). Система уравнений $A\overline{c} = \overline{f}$ имеет вид

$$\begin{cases} 2hc_0 = \dfrac{h}{3}\,(F_0 + 4F_1 + F_2)\,; \\ \dfrac{2h^3}{3}c_1 = \dfrac{h^2}{3}\,(F_2 - F_0)\,. \end{cases}$$

Отсюда находим

$$\begin{cases} c_0 = \dfrac{F_0 + 4F_1 + F_2}{6}; \\ c_1 = \dfrac{F_2 - F_0}{2h}. \end{cases}$$

(5). Элементом наилучшего приближения является функция

$$\varphi(x) = \frac{F_0 + 4F_1 + F_2}{6} + \frac{F_2 - F_0}{2h}(x - x_1).$$

Квадрат отклонения построенного элемента наилучшего приближения от приближаемой функции равен

$$\|f-\varphi\|_{L_2}^2 = \|f\|_{L_2}^2 - \|\varphi\|_{L_2}^2 = \int\limits_{x_1-h}^{x_1+h} f^2(x)\,dx - \int\limits_{x_1-h}^{x_1+h} \varphi^2(x)\,dx =$$

$$= \left\{ \int\limits_a^b G(x)\,dx \approx \frac{b-a}{6}\left(G_0+4G_1+G_2\right) \right\} \approx \frac{h}{3}\left(F_0^2+4F_1^2+F_2^2\right) -$$

$$- \int\limits_{x_1-h}^{x_1+h} \left(\frac{F_0+4F_1+F_2}{6} + \frac{F_2-F_0}{2h}(x-x_1) \right)^2 dx = \frac{h}{3}\left(F_0^2+4F_1^2+F_2^2\right) -$$

$$- \frac{h}{9}\left(2F_0^2+8F_1^2+2F_2^2-2F_0F_2+4F_0F_1+4F_1F_2\right) =$$

$$= \frac{h}{9}\left(F_0-2F_1+F_2\right)^2 = \frac{h^5}{9} f_{\bar{x}x}^2(x_1).$$

Глава 5

Численное решение задачи Коши для обыкновенного дифференциального уравнения

Пусть задана задача Коши для обыкновенного дифференциального уравнения

$$\begin{cases} \dfrac{du(t)}{dt} &= f(t,\, u(t)), \quad 0 < t; \\ u(0) &= u_0. \end{cases}$$

Будем предполагать, что решение задачи Коши существует и единственно. Рассмотрим методы численного решения этой задачи на отрезке $t \in [0;\, T]$.

Введем на сегменте $[0;\, T]$ набор точек

$\omega_\tau = \left\{ t_n = n\tau,\ \tau = \dfrac{T}{N},\ n = 0,\, 1,\, \ldots,\, N \right\}$, который назовём дискретной сеткой. Точки t_n —узлы дискретной сетки. Параметр $\tau > 0$, равный расстоянию между соседними узлами, назовем шагом дискретной сетки. Значение функции $u(t)$, являющейся точным решением задачи Коши, в узле t_n обозначим через $u_n = u(t_n)$. Приближенное значение для u_n, которое вычислено с использованием какого-либо численного метода, обозначим через y_n.

Разность $z_n = y_n - u_n$ назовем погрешностью приближенного решения y_n в узле t_n.

Введём понятие сходимости приближенного решения к точному. Фиксируем точку t_n и построим последовательность дискретных сеток $\{\omega_\tau\}$ такую, что точка t_n является узлом для каждой из дискретных сеток последовательности $\{\omega_\tau\}$. Шаг дискретных сеток у элементов в последовательности $\{\omega_\tau\}$

монотонно стремится к нулю ($\tau \longrightarrow 0$). На каждой из этих дискретных сеток вычисляется приближенное решение y_n.

Определение. Приближенное решение y_n сходится к точному решению u_n в узле t_n, если $|z_n| \stackrel{\tau \to 0}{\longrightarrow} 0$.

Определение. Пусть $|z_n| = O(\tau^p)$, где $p > 0$. Тогда приближенное решение y_n имеет p-й порядок точности в узле t_n.

На сегменте $t \in [t_n;\, t_{n+1}]$ производная $u'(t)$ может быть записана в виде $u'(t) = \dfrac{u_{n+1} - u_n}{\tau} + O(\tau)$.

Заменим исходное дифференциальное уравнение в узле t_n алгебраическим уравнением

$$\frac{y_{n+1} - y_n}{\tau} = f(t_n,\, y_n), \quad n = 0,\, 1,\, \ldots, (N-1)\ .$$

Дополнив его начальным условием $y_0 = u_0$, получим явное разностное уравнение, которое позволяет вычислить все y_n по рекурентной формуле

$$\begin{cases} y_{n+1} &= y_n + \tau f(t_n,\, y_n), \quad n = 0,\, 1,\, \ldots, (N-1); \\ y_0 &= u_0. \end{cases}$$

Возможен и другой вариант при замене дифференциального уравнения в узле t_{n+1} алгебраическим уравнением

$$\frac{y_{n+1} - y_n}{\tau} = f(t_{n+1}, y_{n+1}), \quad n = 0,\, 1,\, \ldots, (N-1)\ .$$

В этом случае появляется неявное разностное уравнение

$$\begin{cases} y_{n+1} - \tau f(t_{n+1}, y_{n+1}) &= y_n, \quad n = 0,\, 1,\, \ldots, (N-1); \\ y_0 &= u_0. \end{cases}$$

Для определения y_{n+1} необходимо решать нелинейное алгебраическое уравнение при каждом $n = 0,\, 1,\, \ldots, (N-1)$.

Явное или неявное разностное уравнение используется для вычисления y_n $(n = 1,\, 2,\, \ldots, (N-1))$, которое принимается за приближённое решение задачи Коши. Приближенное решение y_n может не совпадать с значением точного решения u_n в узле t_n. Поэтому, формальная подстановка в разностное уравнение точного решения задачи Коши может нарушать равенство в разностном уравнении. Возникающий дисбаланс в разностном уравнении называется погрешностью аппроксимации (невязкой) исходного дифференциального уравнения разностным уравнением. Погрешность аппроксимации в узле

t_n будем обозначать через ψ_n. Например, для явного разностного уравнения погрешность аппроксимации в узле t_n равна $\psi_n = -\dfrac{u_{n+1} - u_n}{\tau} + f(t_n, u_n)$.

Определение. Разностное уравнение аппроксимирует исходное дифференциальное уравнение в узле t_n, если $|\psi_n| \xrightarrow{\tau \to 0} 0$.

Определение. Разностное уравнение имеет p-й порядок аппроксимации в узле t_n, если $|\psi_n| = O(\tau^p)$.

5.1 Методы Рунге-Кутта

Методами Рунге-Кутта называется семейство разностных уравнений вида:

$$\frac{y_{n+1} - y_n}{\tau} = \sum_{i=1}^{m} \sigma_i K_i(y), \quad n = 0, 1, \ldots, y_0 = u_0, \tag{5.1}$$

где величины $K_i(y)$ вычисляются по следующим формулам:

$$\begin{aligned}
K_1(y) &= f(t_n, y_n); \\
K_2(y) &= f(t_n + \tau a_2, y_n + \tau b_{21} K_1(y)); \\
K_3(y) &= f(t_n + \tau a_3, y_n + \tau b_{31} K_1(y) + \tau b_{32} K_2(y)); \\
&\ldots \\
K_m(y) &= f(t_n + \tau a_m, y_n + \tau \sum_{i=1}^{m-1} b_{mi} K_i(y)).
\end{aligned}$$

Выбор числовых значений параметров m, σ_i, a_i, b_{ij} задает конкретный метод Рунге-Кутта.

Разностное уравнение (5.1) используется для определения y_{n+1}. Особенностью методов Рунге-Кутта является то, что при каждом $n = 0, 1, \ldots$ необходимо m раз вычислить функцию f от различных значений аргументов. Поэтому, эти методы принято называть m-этапными методами.

Сходимость y_n $(n = 1, 2, \ldots)$ в узлах дискретной сетки к значению точного решения задачи Коши обусловлена следующей теоремой.

Теорема 5.1 (О сходимости методов Рунге-Кутта). Пусть метод Рунге-Кутта аппроксимирует заданное обыкновенное дифференциальное уравнение. Тогда приближенное решение y_n сходится к точному решению u_n и порядок точности приближенного решения совпадает с порядком аппроксимации разностным уравнением обыкновенного дифференциального уравнения.

Доказательство. Будем предполагать, что функция $f(t, u)$ Липшиц непрерывна по второму аргументу с константой L, то есть для любых u_1, u_2 верно неравенство $|f(t, u_1) - f(t, u_2)| \leqslant L|u_1 - u_2|$.

Погрешность приближенного решения равна $z_n = y_n - u_n$. Представим y_n в виде $y_n = z_n + u_n$ и подставим в (5.1). В результате получим

$$\frac{z_{n+1} - z_n}{\tau} = -\frac{u_{n+1} - u_n}{\tau} + \sum_{i=1}^{m} \sigma_i K_i(u) + \sum_{i=1}^{m} \sigma_i (K_i(y) - K_i(u)).$$

Заметим, что комбинация $-\dfrac{u_{n+1} - u_n}{\tau} + \sum\limits_{i=1}^{m} \sigma_i K_i(u) = \psi_n$ является погрешностью аппроксимации разностным уравнением исходного обыкновенного дифференциального уравнения. Введём обозначение $\overline{\psi_n} = \sum\limits_{i=1}^{m} \sigma_i (K_i(y) - K_i(u))$. Тогда (5.1) примет вид

$$\frac{z_{n+1} - z_n}{\tau} = \psi_n - \overline{\psi_n}.$$

Получим оценку на $\left|\overline{\psi_n}\right|$. Сначала рассмотрим выражение $|K_i(y) - K_i(u)|$ для значений $i = 1, 2, 3$.

$\mathbf{i = 1}$: $|K_1(y) - K_1(u)| = |f(t_n, y_n) - f(t_n, u_n)| \leqslant L|y_n - u_n| = L|z_n|$.

$\mathbf{i = 2}$: $|K_2(y) - K_2(u)| = |f(t_n + \tau a_2, y_n + \tau b_{21} K_1(y)) - f(t_n + \tau a_2, u_n + \tau b_{21} K_1(u))| \leqslant L|y_n - u_n + \tau b_{21}(K_1(y) - K_1(u))| \leqslant L(|z_n| + \tau |b_{21}||K_1(y) - K_1(u)|) \leqslant L|z_n|(1 + \tau bL)$,

где $b = \max\limits_{\substack{i=\overline{2,m} \\ j=\overline{1,(i-1)}}} |b_{ij}|$.

$\mathbf{i = 3}$: $|K_3(y) - K_3(u)| = |f(t_n + \tau a_3, y_n + \tau b_{31} K_1(y) + \tau b_{32} K_2(y)) - f(t_n + \tau a_3, u_n + \tau b_{31} K_1(u) + \tau b_{32} K_2(u))| \leqslant L|y_n - u_n + \tau b_{31}(K_1(y) - K_1(u)) + \tau b_{32}(K_2(y) - K_2(u))| \leqslant L(|z_n| + \tau |b_{31}||K_1(y) - K_1(u)| + \tau |b_{32}| \cdot |K_2(y) - K_2(u)|) \leqslant L(|z_n| + \tau bL|z_n| + \tau bL|z_n|(1 + \tau bL)) = L|z_n|(1 + \tau bL)^2$.

Неравенство $|K_i(y) - K_i(u)| \leqslant L|z_n|(1 + \tau bL)^{i-1}$ верно при $i = 1, 2, 3$. Пусть это неравенство выполняется для индекса $i \geq 3$. Покажем, что оно верно и для следующего индекса $(i + 1)$:

$$|K_{i+1}(y) - K_{i+1}(u)| = \left| f\left(t_n + \tau a_{i+1}, y_n + \tau \sum_{j=1}^{i} b_{(i+1)j} K_j(y)\right) - \right.$$

$$-f\left(t_n+\tau a_{i+1},\, u_n+\tau\sum_{j=1}^{i}b_{(i+1)j}K_j(u)\right)\Bigg| \leqslant$$

$$\leqslant L\left|y_n-u_n+\tau\left(\sum_{j=1}^{i}b_{(i+1)j}(K_j(y)-K_j(u))\right)\right| \leqslant$$

$$\leqslant L\left(|z_n|+\tau b\sum_{j=1}^{i}|K_j(y)-K_j(u)|\right) \leqslant L\left(|z_n|+\tau bL|z_n|\cdot\right.$$

$$\left.\cdot\sum_{j=1}^{i}(1+\tau bL)^{j-1}\right) = L|z_n|\left(1+\tau bL\frac{1-(1+\tau bL)^i}{1-(1+\tau bL)}\right) = L|z_n|(1+\tau bL)^i.$$

Теперь получим оценку на $|\overline{\psi_n}|$:

$$|\overline{\psi_n}| \leqslant \sum_{i=1}^{m}|\sigma_i||K_i(y)-K_i(u)| \leqslant |z_n|L\sigma\sum_{i=1}^{m}(1+\tau bL)^{i-1} \leqslant |z_n|L\sigma(1+\tau bL)^{m-1}m,$$

где $\sigma = \max\limits_{i=\overline{1,m}}|\sigma_i|$.

Так как

$$(1+\tau bL)^{m-1} \leqslant \{(1+\mathsf{x})^{\mathsf{s}} \leqslant \mathsf{e}^{\mathsf{sx}}\} \leqslant e^{\tau bL(m-1)} \leqslant \{\tau \leqslant \mathsf{T}\} \leqslant e^{TbL(m-1)},$$

то получаем равномерную по τ оценку $|\overline{\psi_n}| \leqslant |z_n|L\sigma\alpha$, где $\alpha = me^{TbL(m-1)}$.

Вернёмся к уравнению для погрешности приближённого решения $\dfrac{z_{n+1}-z_n}{\tau} = \psi_n - \overline{\psi_n}$, из которого следует неравенство

$$|z_{n+1}| \leqslant |z_n|+\tau|\psi_n|+\tau|\overline{\psi_n}| \leqslant |z_n|+\tau|\psi_n|+\tau|z_n|\sigma L\alpha = |z_n|(1+\tau\sigma L\alpha)+ \\ +\tau|\psi_n|,$$

верное для любого $n = 0,\, 1,\, \ldots$.

Используем это неравенство при получении оценки на $|z_n|$:

$$\begin{aligned}
|z_n| &\leqslant |z_{n-1}|(1+\tau\sigma L\alpha)+\tau|\psi_{n-1}| \leqslant \\
&\leqslant (|z_{n-2}|(1+\tau\sigma L\alpha)+\tau|\psi_{n-2}|)\,(1+\tau\sigma L\alpha)+\tau|\psi_{n-1}| = \\
&= |z_{n-2}|(1+\tau\sigma L\alpha)^2+\tau|\psi_{n-2}|(1+\tau\sigma L\alpha)+\tau|\psi_{n-1}| \leqslant \ldots \leqslant \\
&\leqslant |z_0|(1+\tau\sigma L\alpha)^n+\tau\sum_{j=0}^{n-1}|\psi_{n-1-j}|(1+\tau\sigma L\alpha)^j = \\
&= \{z_0 = y_0-u_0 = 0\} = \tau\sum_{j=0}^{n-1}|\psi_{n-1-j}|(1+\tau\sigma L\alpha)^j.
\end{aligned}$$

Введём обозначение $\Psi = \max\limits_{j=\overline{0,n-1}} |\psi_j|$. Тогда

$$|z_n| \leqslant \Psi\tau \sum_{j=0}^{n-1} (1+\tau\sigma L\alpha)^j \leqslant \Psi\tau n(1+\tau\sigma L\alpha)^{n-1} \leqslant \Psi T e^{T\sigma L\alpha},$$

так как $\tau n \leqslant T$.

Отсюда следует, что если разностное уравнение аппроксимирует во всех узлах дискретной сетки обыкновенное дифференциальное уравнение, то есть $|\Psi| \xrightarrow{\tau\to 0} 0$, то приближённое решение y_n сходится к точному решению u_n, то есть $|z_n| \xrightarrow{\tau\to 0} 0$. Кроме того, если разностное уравнение аппроксимирует обыкновенное дифференциальное уравнение с p-м порядком $(|\Psi| = O(\tau^p))$, то приближённое решение имеет p-й порядок точности $(|z_n| = O(\tau^p))$.

Теорема доказана. □

Двухэтапные методы Рунге-Кутта второго порядка аппроксимации

Для определения числовых значений параметров методов Рунге-Кутта конструктивным является условие, что m-этапный метод Рунге-Кутта имеет m-ый порядок аппроксимации.

Рассмотрим двухэтапные методы Рунге-Кутта ($m = 2$). Разностное уравнение для вычисления приближенного решения y_{n+1} имеет вид:

$$\begin{cases} \dfrac{y_{n+1}-y_n}{\tau} = \sigma_1 K_1(y) + \sigma_2 K_2(y), \quad n = 0,\, 1,\, \ldots, \quad y_0 = u_0; \\ K_1(y) = f(t_n,\, y_n); \\ K_2(y) = f(t_n + \tau a_2,\, y_n + \tau b_{21}\, K_1(y)). \end{cases}$$

Погрешность аппроксимации для двухэтапных методов Рунге-Кутта равна

$$\psi_n = -\frac{u_{n+1}-u_n}{\tau} + \sigma_1 K_1(u) + \sigma_2 K_2(u).$$

Выберем параметры σ_1, σ_2, a_2 и b_{21} так, чтобы двухэтапный метод Рунге-Кутта имел бы второй порядок аппроксимации, то есть $\psi_n = O(\tau^2)$.

Проведем ряд промежуточных преобразований. Запишем разностную производную в виде

$$\frac{u_{n+1}-u_n}{\tau} = \frac{1}{\tau}(u(t_n+\tau) - u_n) = \frac{1}{\tau}(u_n + u'_n\tau + u''_n\frac{\tau^2}{2} + O(\tau^3) - u_n) = \\ = u'_n + u''_n\frac{\tau}{2} + O(\tau^2).$$

Так как $u' = f(t,u)$, то, предполагая достаточную гладкость у функций $u(t)$ и $f(t,u)$, получим, что $u'' = f_t(t,u) + f_u(t,u)u' = f_t(t,u) + f_u(t,u)f(t,u)$.

Тогда

$$\frac{u_{n+1}-u_n}{\tau} = f(t_n,u_n) + \left(f_t(t_n,u_n) + f_u(t_n,u_n)f(t_n,u_n)\right)\frac{\tau}{2} + O(\tau^2).$$

Функция $K_1(u) = f(t_n,\, u_n)$, а для $K_2(u)$ верно представление

$$K_2(u) = f(t_n + \tau a_2,\, u_n + \tau b_{21} f(t_n,u_n)) = f(t_n,u_n) + f_t(t_n,\, u_n)\tau a_2 + \\ + f_u(t_n,u_n)f(t_n,u_n)\tau b_{21} + O(\tau^2).$$

С учетом предыдущих преобразований, выражение для ψ_n принимает вид

$$\psi_n = -f(t_n,u_n) - \left(f_t(t_n,u_n) + f_u(t_n,u_n)f(t_n,u_n)\right)\frac{\tau}{2} + \sigma_1 f(t_n,u_n) + \\ + \sigma_2(f(t_n,u_n) + f_t(t_n,\, u_n)a_2\tau + f_u(t_n,\, u_n)f(t_n,u_n)\tau b_{21}) + O(\tau^2) = \\ = f(t_n,u_n)(-1+\sigma_1+\sigma_2) + f_t(t_n,\, u_n)(-\frac{1}{2} + \sigma_2 a_2)\tau + \\ + f_u(t_n,\, u_n)f(t_n,u_n)(-\frac{1}{2} + \sigma_2 b_{21})\tau + O(\tau^2).$$

Для того, чтобы $\psi_n = O(\tau^2)$, потребуем выполнения системы уравнений

$$\begin{cases} \sigma_1 + \sigma_2 = 1; \\ \sigma_2 a_2 = \dfrac{1}{2}; \\ \sigma_2 b_{21} = \dfrac{1}{2}. \end{cases}$$

Из этих уравнений следует, что $a_2 = b_{21}$. Введём обозначения $\sigma_2 = \sigma$ и $a_2 = a$. Тогда предыдущая система уравнений принимает вид

$$\begin{cases} \sigma_1 = 1 - \sigma; \\ \sigma a = \dfrac{1}{2}. \end{cases}$$

Следовательно, разностное уравнение

$$y_{n+1} - y_n = \tau((1-\sigma)f(t_n,\, y_n) + \sigma f(t_n + a\tau,\, y_n + \tau a f(t_n,\, y_n)))$$

имеет 2-й порядок аппроксимации и приближённое решение

$y_n\,(n = 1,\, 2,\, \ldots)$ имеет 2-й порядок точности, если параметры σ и a удовлетворяют условию $\sigma a = \dfrac{1}{2}$.

Рассмотрим два частных случая. Пусть $\sigma = 1$, $a = \dfrac{1}{2}$. Тогда разностное уравнение принимает вид

$$y_{n+1} - y_n = \tau f(t_n + \frac{\tau}{2},\, y_n + \frac{\tau}{2} f(t_n,\, y_n)).$$

Выбору $\sigma = \dfrac{1}{2}$, $a = 1$ соответствует разностное уравнение

$$y_{n+1} - y_n = \frac{\tau}{2}\left(f(t_n,\, y_n) + f(t_n + \tau,\, y_n + \tau f(t_n,\, y_n))\right).$$

Графическая иллюстрация разностного уравнения $\dfrac{y_{n+1} - y_n}{\tau} = \sum\limits_{i=1}^{m} \sigma_i K_i(y)$, которое используется в методах Рунге-Кутта для определения y_{n+1}, представлена на рис.5.1. Пусть координаты точки в плоскости задаются переменными t и $u(t)$. Построим прямоугольный треугольник ΔABC, у которого точка A имеет координаты $(t_n,\, y_n)$, точка B имеет координаты $(t_{n+1},\, y_n)$ и угол α задаётся из условия $tg\,\alpha = \sum\limits_{i=1}^{m} \sigma_i K_i(y)$.

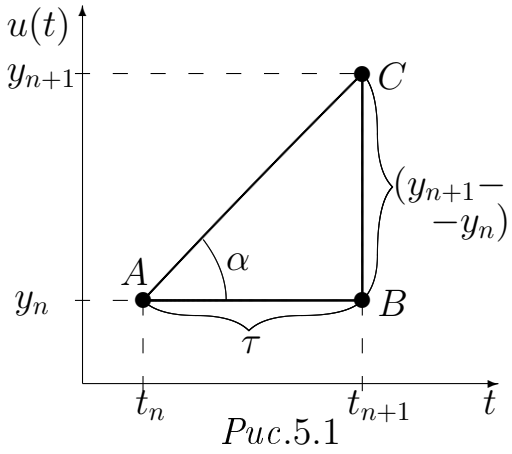

*Рис.*5.1

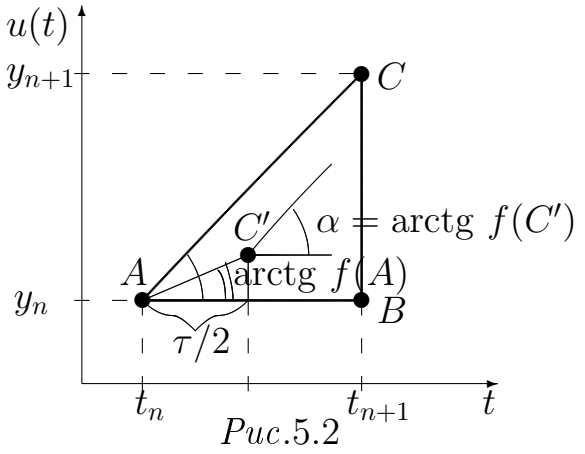

*Рис.*5.2

Тогда, точка C будет иметь координаты $(t_{n+1},\, y_{n+1})$.

В случае двухэтапного, второго порядка аппроксимации метода Рунге-Кутта с параметрами $\sigma = 1$ и $a = \dfrac{1}{2}$ угол α определяется из условия, что $tg\,\alpha$ равен значению функции f от аргументов, соответствующих координатам точки $C'\left(t_n + \dfrac{\tau}{2},\, y_n + \dfrac{\tau}{2} f(t_n,\, y_n)\right)$ (рис.5.2), то есть $tg\,\alpha = f(C')$.

Если выбраны параметры $\sigma = \dfrac{1}{2}$ и $a = 1$, то $tg\,\alpha$ равен полусумме значений функции f от аргументов, соответствующих значениям координат точки $A\,(t_n,\, y_n)$ и точки $C''\,(t_{n+1},\, y_n + \tau f(t_n,\, y_n))$ (рис.5.3), то есть $tg\,\alpha = \dfrac{1}{2}\left(f(A) + f(C'')\right)$.

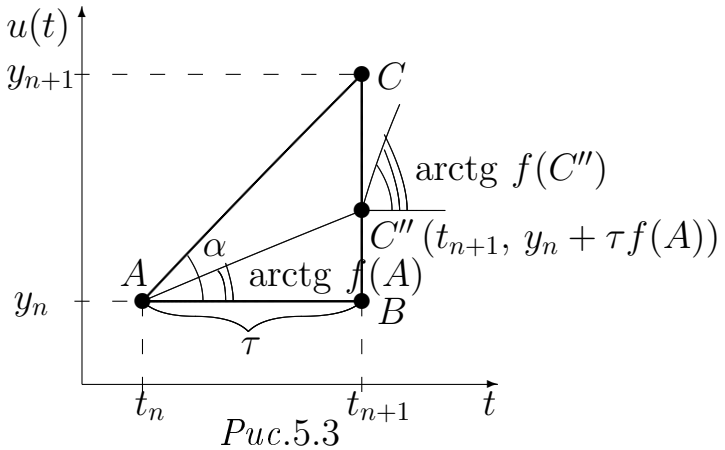

*Рис.*5.3

Методы Рунге-Кутта четвертого порядка точности

Методы Рунге-Кутта с параметром $m \geq 5$ в вычислительной практике используются редко. Наиболее часто применяются разностные уравнения соответствующие 4-х этапным методам Рунге-Кутта 4-ого порядка аппроксимации. В качестве примера приведем одно из таких разностных уравнений, обладающее свойством вычислительной устойчивости по отношению к ошибкам машинного округления чисел. Это разностное уравнение, используемое в стандартных программах, имеет вид :

$$\begin{cases} \dfrac{y_{n+1}-y_n}{\tau} = \dfrac{1}{6}(K_1(y)+2K_2(y)+2K_3(y)+K_4(y)),\ n=0,\,1,\,\ldots,\ y_0=u_0; \\ K_1(y) = f(t_n,\,y_n); \\ K_2(y) = f(t_n+\dfrac{\tau}{2},\,y_n+\dfrac{\tau}{2}K_1(y)); \\ K_3(y) = f(t_n+\dfrac{\tau}{2},\,y_n+\dfrac{\tau}{2}K_2(y)); \\ K_4(y) = f(t_n+\tau,\,y_n+\tau K_3(y)). \end{cases}$$

5.2 Многошаговые методы

При реализации методов Рунге-Кутта необходимо вычислять функцию $f(t,\,u)$ от произвольных значений её аргументов. Это возможно, например, при явном задании функции формулой.

Рассмотрим методы, в которых значения функции $f(t,\,u)$ достаточно знать лишь в фиксированных узлах дискретной сетки.

Определение. Будем называть m-шаговым методом расчетную схему следующего вида

$$\frac{a_0 y_n + a_1 y_{n-1} + \ldots + a_m y_{n-m}}{\tau} = b_0 f_n + b_1 f_{n-1} + \ldots + b_m f_{n-m},\ n \geq m, \quad (5.2)$$

где m, a_i, b_i $(i = \overline{0,\ m})$ — числовые параметры метода, а $y_{n-i} = y(t_{n-i})$ и $f_{n-i} = f(t_{n-i},\ y_{n-i})$.

Соотношение (5.2) является уравнением относительно y_n. Для начала расчёта по схеме (5.2) необходимо знать значения y_0, y_1, …, y_{m-1}. Значение $y_0 = u_0$ — начальное условие дифференциальной задачи. Остальные начальные данные, а именно y_1, …, y_{m-1} необходимо подготовить используя другие численные методы, например, методы Рунге-Кутта соответствующего порядка точности.

В случае, когда в (5.2) коэффициент b_0 равен нулю, то есть в правой части зависимость от y_n отсутствует, соответствующий метод называется явным. Если $b_0 \neq 0$, то (5.2) представляет собой нелинейное уравнение относительно y_n, которое нужно решать, используя, например, итерационный метод Нютона. В этом случае метод (5.2) называется неявным.

Заметим, что если уравнение (5.2) умножить на какую-нибудь ненулевую константу, то y_n не изменится. Устраним неоднозначность значений коэффициентов a_i и b_i. Для этого введём условие нормировки $\sum\limits_{i=0}^{m} b_i = 1$.

Покажем, что при выполнении этого условия правая часть уравнения (5.2) аппроксимирует правую часть дифференциального уравнения. Рассмотрим разность правой части дифференциального уравнения и правой части (5.2) на точном решении дифференциальной задачи:

$$f(t_n, u_n) - \sum_{i=0}^{m} b_i f(t_n - i\tau,\ u(t_n - i\tau)) = f(t_n, u_n) - \sum_{i=0}^{m} b_i \left(f(t_n, u_n) + O(\tau)\right) =$$

$$= f_n \cdot \left(1 - \sum_{i=0}^{m} b_i\right) + O(\tau) = O(\tau)$$

Теперь получим достаточные условия для «k-ого» порядка аппроксимации расчётной схемой (5.2) дифференциального уравнения. Рассмотрим выражение для невязки:

$$\psi_n = -\frac{1}{\tau} \sum_{i=0}^{m} a_i u_{n-i} + \sum_{i=0}^{m} b_i f(t_{n-i},\ u_{n-i}).$$

Верны следующие разложения в ряд Тейлора:

$$u_{n-i} = u(t_n - i\tau) = \sum_{j=0}^{k} \frac{u_n^{(j)}}{j!}(-i\tau)^j + O(\tau^{k+1});$$

$$f(t_{n-i},\, u_{n-i}) = \{\mathsf{u}' = \mathsf{f}(\mathsf{t},\mathsf{u})\} = u'_{n-i} = u'(t_n - i\tau) = \sum_{j=0}^{k-1} \frac{u_n^{(j+1)}}{j!}(-i\tau)^j + O(\tau^k).$$

Подставив эти разложения в выражение для невязки, получим

$$\psi_n = -\frac{1}{\tau}\sum_{i=0}^{m} a_i \sum_{j=0}^{k}(-1)^j \frac{u_n^{(j)}}{j!} i^j \tau^j + \sum_{i=0}^{m} b_i \sum_{j=0}^{k-1}(-1)^j \frac{u_n^{j+1}}{j!} i^j \tau^j + O(\tau^k).$$

Изменим в двойных суммах порядок суммирования. После этого, из первой двойной суммы выделим слагаемое с $j = 0$, а во второй двойной сумме сделаем замену индекса суммирования $j = j'-1$. Тогда выражение для невязки примет вид:

$$\begin{aligned}
\psi_n &= -\frac{1}{\tau}\sum_{i=0}^{m} u_n a_i - \frac{1}{\tau}\sum_{j=1}^{k}(-1)^j \frac{u_n^{(j)}}{j!}\tau^j \sum_{i=0}^{m} a_i i^j + \\
&+ \sum_{j'=1}^{k}(-1)^{j'-1}\frac{u_n^{(j')}}{(j'-1)!}\tau^{(j'-1)}\sum_{i=0}^{m} b_i i^{j'-1} + O(\tau^k) = \\
&= -\frac{u_n}{\tau}\sum_{i=0}^{m} a_i + \sum_{j=1}^{k}(-1)^{j-1}\frac{u_n^{(j)}}{j!}\tau^{j-1}\sum_{i=0}^{m} i^{j-1}(i a_i + j b_i) + O(\tau^k).
\end{aligned}$$

Следовательно, для того чтобы $\psi_n = O(\tau^k)$, достаточно потребовать выполнения системы равенств:

$$\begin{cases}
\sum\limits_{i=0}^{m} a_i = 0; \\
\sum\limits_{i=0}^{m} i^{j-1}(i a_i + j b_i) = 0, \quad j = \overline{1,\, k}.
\end{cases} \tag{5.3}$$

Второе уравнение в (5.3) при $j = 1$ имеет вид $\sum\limits_{i=0}^{m} i a_i + \sum\limits_{i=0}^{m} b_i = 0$. Так как $\sum\limits_{i=0}^{m} b_i = 1$, то $\sum\limits_{i=0}^{m} i a_i = -1$. Первое слагаемое в этой сумме равно нулю. Поэтому $\sum\limits_{i=1}^{m} i a_i = -1$.

Объединяя вместе последнее равенство, условие нормировки и (5.3), получаем следующую систему условий, обеспечивающих k-ый порядок аппроксимации:

$$\begin{cases} \sum\limits_{i=0}^{m} b_i = 1; \\ \sum\limits_{i=0}^{m} a_i = 0; \\ \sum\limits_{i=1}^{m} i a_i = -1; \\ \sum\limits_{i=1}^{m} i^{j-1}(i a_i + j b_i) = 0, \quad j = \overline{2,\, k}. \end{cases} \tag{5.4}$$

Равенства (5.4) представляют собой систему из $(k+2)$ линейных алгебраических уравнений относительно неизвестных a_i, b_i $(i = \overline{0,\, m})$. Количество неизвестных равно $2m+2$. Система (5.4) имеет решение (возможно не единственное), если выполнено неравенство $k + 2 \leqslant 2m + 2$.

Таким образом, k — порядок аппроксимации m-шагового метода не может превышать значения $2m$ (для неявного метода). Если выполнено дополнительное условие $b_0 = 0$, то в системе (5.4) неизвестных на единицу меньше, и максимально возможный порядок аппроксимации будет равен $(2m-1)$.

5.3 Методы Адамса и Гира

В вычислительной практике преимущественно используют частные случаи m-шаговых методов. Перейдём к их рассмотрению.

5.3.1 Многошаговые методы Адамса

Определение. Методами Адамса называют m-шаговые методы, в которых $a_0 = 1,\ a_1 = -1,\ a_2 = a_3 = \ldots = a_m = 0$.

Таким образом, в методах Адамса расчётная схема для вычисления y_n имеет вид

$$\frac{y_n - y_{n-1}}{\tau} = b_0 f_n + b_1 f_{n-1} + \ldots + b_m f_{n-m},\ n \geq m. \tag{5.5}$$

Подставив значения a_i, определяющие методы Адамса, в (5.4), получим следующую систему условий, обеспечивающих k-ый порядок аппроксимации

для методов Адамса:

$$\begin{cases} \sum\limits_{i=0}^{m} b_i = 1; \\ j \sum\limits_{i=1}^{m} i^{j-1} b_i = 1, \quad j = \overline{2,\, k}. \end{cases} \tag{5.6}$$

Система (5.6) является системой линейных алгебраических уравнений относительно b_i ($i = \overline{0,\, m}$). Система (5.6) состоит из k уравнений и содержит $(m+1)$ неизвестных. Система не является переопределённой, если $k \leqslant m+1$.

Таким образом, порядок аппроксимации в неявных методах Адамса не может превышать значения $(m+1)$. В явных методах Адамса ($b_0 = 0$) порядок аппроксимации не может быть больше m ($k \leqslant m$).

Рассмотрим некоторые примеры.

Пример 5.1. Метод Адамса явный ($b_0 = 0$), одношаговый ($m = 1$), максимального порядка аппроксимации ($k = 1$).

Достаточные условия k-ого порядка аппроксимации, то есть система (5.6), в данном случае имеет вид $b_1 = 1$.

Тогда, расчетная схема (5.5) записывается следующим образом:

$$\frac{y_n - y_{n-1}}{\tau} = f(t_{n-1},\, y_{n-1}), \quad n \geqslant 1.$$

Для начала вычислений необходимо одно начальное условие $y_0 = u_0$.

Пример 5.2. Метод Адамса явный ($b_0 = 0$), трехшаговый ($m = 3$), максимального порядка аппроксимации ($k = 3$).

Система уравнений (5.6) имеет следующий вид:

$$\begin{cases} b_1 + b_2 + b_3 = 1; \\ 2(b_1 + 2b_2 + 3b_3) = 1; \\ 3(b_1 + 4b_2 + 9b_3) = 1; \end{cases} \iff \begin{cases} b_1 = \dfrac{23}{12}; \\ b_2 = -\dfrac{4}{3}; \\ b_3 = \dfrac{5}{12}. \end{cases}$$

Расчетная схема этого метода такова:

$$\frac{y_n - y_{n-1}}{\tau} = \frac{23 f_{n-1} - 16 f_{n-2} + 5 f_{n-3}}{12}, \; n \geqslant 3.$$

Для начала расчёта нужны дополнительные данные y_0, y_1 и y_2. Для всех расчетных схем $y_0 = u_0$. Значения y_1 и y_2 рассчитываются с использованием

методов, например, Рунге-Кутта, которые должны быть не менее, чем третьего порядка точности, так как $k = 3$. Ошибка в начальных данных не должна быть больше погрешности расчетной схемы.

Пример 5.3. Метод Адамса неявный ($b_0 \neq 0$), одношаговый ($m = 1$), максимального порядка аппроксимации ($k = 2$).

Система уравнений (5.6) в данном случае принимает вид:

$$\begin{cases} b_0 + b_1 = 1; \\ 2b_1 = 1; \end{cases} \iff \begin{cases} b_1 = \dfrac{1}{2}; \\ b_2 = \dfrac{1}{2}. \end{cases}$$

Тогда, расчётная схема метода имеет вид:

$$\frac{y_n - y_{n-1}}{\tau} = \frac{f_n + f_{n-1}}{2}, \quad n \geqslant 1.$$

Для начала вычислений необходимо одно начальное условие $y_0 = u_0$.

При каждом $n \geqslant 1$, для нахождения y_n нужно решать нелинейное алгебраическое уравнение

$$y_n - \frac{\tau}{2} f(t_n,\, y_n) = y_{n-1} + \frac{\tau}{2} f_{n-1}.$$

Эффективным методом решения этого уравнения является итерационный метод Ньютона. Пусть

$$F(x) = x - \frac{\tau}{2} f(t_n,\, x) - y_{n-1} - \frac{\tau}{2} f_{n-1}.$$

Тогда, искомое y_n является корнем уравнения $F(x) = 0$. Для приближённого вычисления значения этого корня строится последовательность итерационных приближений

$$x^{s+1} = x^s - \frac{F(x^s)}{F'(x^s)},\ s = 0,\, 1, \ldots,$$

где $F'(x^s) = 1 - \dfrac{\tau}{2} f_u(t_n,\, x^s)$ и $x^0 = y_{n-1}$.

Пример 5.4. Метод Адамса неявный ($b_0 \neq 0$), двухшаговый ($m = 2$), максимального порядка аппроксимации ($k = 3$).

Система уравнений (5.6) в данном случае такова:

$$\begin{cases} b_0 + b_1 + b_2 = 1; \\ 2(b_1 + 2b_2) = 1; \\ 3(b_1 + 4b_2) = 1; \end{cases} \iff \begin{cases} b_0 = \dfrac{5}{12}; \\ b_1 = \dfrac{2}{3}; \\ b_2 = -\dfrac{1}{12}. \end{cases}$$

Расчетная схема метода имеет вид:

$$\frac{y_n - y_{n-1}}{\tau} = \frac{5f_n + 8f_{n-1} - f_{n-2}}{12}, \quad n \geqslant 2.$$

Для вычислений по этой схеме нужно знать y_0 и y_1. Начальное условие $y_0 = u_0$, а y_1 рассчитывается, например, по методу Рунге-Кутта 3-го порядка точности.

При каждом $n \geqslant 2$ значение y_n является корнем нелинейного алгебраического уравнения, которое решается с использованием итерационного метода Ньютона.

5.3.2 Многошаговые методы Гира

Определение. Методами Гира называют m-шаговые методы, в которых $b_0 = 1,\ b_1 = b_2 = \ldots = b_m = 0$.

В методах Гира расчетная схема для вычисления y_n имеет вид

$$\frac{a_0 y_n + a_1 y_{n-1} + \ldots + a_m y_{n-m}}{\tau} = f_n,\ n \geq m.$$

Отметим, что все методы Гира являются неявными и для любого $n \geq m$ значение y_n находится в результате решения нелинейного уравнения

$$a_0 y_n - \tau f(t_n,\, y_n) = \sum_{i=1}^{m} a_i y_{n-i}.$$

Подставив значения b_i, определяющие методы Гира, в (5.4), получим следующую систему условий, обеспечивающих k-ый порядок аппроксимации для методов Гира:

$$\begin{cases} \sum\limits_{i=0}^{m} a_i = 0; \\ \sum\limits_{i=1}^{m} i a_i = -1; \\ \sum\limits_{i=1}^{m} i^j a_i = 0, \quad j = \overline{2,\, k}. \end{cases} \tag{5.7}$$

Система (5.7) состоит из $(k+1)$ уравнений относительно $(m+1)$ неизвестных, которыми являются a_i. Система (5.7) не переопределена, если выполняется неравенство $k + 1 \leq m + 1$. То есть порядок аппроксимации k в методах Гира не больше чем m.

Рассмотрим некоторые примеры.

Пример 5.5. Метод Гира одношаговый ($m = 1$), максимального порядка аппроксимации ($k = 1$).

Система уравнений (5.7) в данном случае принимает вид:

$$\begin{cases} a_0 + a_1 = 0; \\ a_1 = -1; \end{cases} \iff \begin{cases} a_0 = 1; \\ a_1 = -1. \end{cases}$$

Расчётная схема метода имеет вид:

$$\frac{y_n - y_{n-1}}{\tau} = f(t_n, y_n), \quad n \geqslant 1.$$

Для вычислений по этой схеме необходимо одно начальное условие $y_0 = u_0$.

При каждом $n \geqslant 1$ значение y_n является корнем нелинейного алгебраического уравнения, которое решается с использованием итерационного метода Ньютона.

Пример 5.6. Метод Гира двухшаговый ($m = 2$), максимального порядка аппроксимации ($k = 2$).

Система уравнений (5.7) имеет следующий вид:

$$\begin{cases} a_0 + a_1 + a_2 = 0; \\ a_1 + 2a_2 = -1; \\ a_1 + 4a_2 = 0; \end{cases} \iff \begin{cases} a_0 = \dfrac{3}{2}; \\ a_1 = -2; \\ a_2 = \dfrac{1}{2}. \end{cases}$$

Расчетная схема этого метода такова:

$$\frac{3y_n - 4y_{n-1} + y_{n-2}}{2\tau} = f(t_n, y_n), \quad n \geqslant 2.$$

Для начала расчёта нужны дополнительные данные $y_0 = u_0$ и y_1. Значение y_1 рассчитывается с использованием методов, например, Рунге-Кутта, которые должны быть не менее, чем второго порядка точности.

При каждом $n \geqslant 2$ значение y_n является корнем нелинейного алгебраического уравнения

$$\frac{3}{2}y_n - \tau f(t_n\, y_n) = 2y_{n-1} - \frac{1}{2}y_{n-2},$$

которое решается с использованием итерационного метода Ньютона.

Пример 5.7. Метод Гира трехшаговый ($m = 3$), максимального порядка аппроксимации ($k = 3$).

Система уравнений (5.7) в данном случае принимает вид:

$$\begin{cases} a_0 + a_1 + a_2 + a_3 = 0; \\ a_1 + 2a_2 + 3a_3 = -1; \\ a_1 + 4a_2 + 9a_3 = 0; \\ a_1 + 8a_2 + 27a_3 = 0; \end{cases} \iff \begin{cases} a_0 = \dfrac{11}{6}; \\ a_1 = -3; \\ a_2 = \dfrac{3}{2}; \\ a_3 = -\dfrac{1}{3}. \end{cases}$$

Расчетная схема метода имеет вид:

$$\frac{11y_n - 18y_{n-1} + 9y_{n-2} - 2y_{n-3}}{6\tau} = f(t_n,\, y_n), \quad n \geqslant 3.$$

Для начала расчета нужны дополнительные данные $y_0 = u_0$, y_1 и y_2. Значения y_1 и y_2 рассчитываются с использованием, например, методов Рунге-Кутта, которые должны быть не менее, чем третьего порядка точности.

При каждом $n \geqslant 3$ значение y_n является корнем нелинейного алгебраического уравнения, которое решается с использованием итерационного метода Ньютона.

Замечание. В вычислительной практике используются методы Гира, имеющие десятый порядок аппроксимации и обладающие свойством вычислительной устойчивости.

5.4 Устойчивость численных методов решения задачи Коши

Рассмотрим задачу Коши для обыкновенного дифференциального уравнения первого порядка следующего вида

$$\begin{cases} \dfrac{du(t)}{dt} = \lambda u, \quad t > 0, \quad \lambda = \text{const} < 0; \\ u(0) = u_0. \end{cases} \tag{5.8}$$

Решение этой задачи имеет вид $u(t) = u_0 \exp(\lambda t)$. При $\lambda < 0$ функция $u(t)$ с ростом t монотонно стремится к нулю, то есть для любого $t \geq 0$ выполняется неравенство $|u(t)| \leq |u_0|$.

В вычислительной практике принято, в силу ряда причин, использовать задачу (5.8) в качестве теста для проверки свойств различных численных методов решения задачи Коши. Должно выполняться условие, что $|y_n|$ — модуль

приближённого численного решения задачи (5.8), полученного с использованием конкретного численного метода, монотонно стремится к нулю с ростом n.

Применим для численного решения задачи (5.8) явный одношаговый метод Адамса (см. Пример 5.1.). Расчётная схема метода имеет вид

$$\frac{y_{n+1} - y_n}{\tau} = f(t_n, y_n),$$

где правая часть $f(t_n, y_n) = \lambda y_n$.

Отсюда следует соотношение $y_{n+1} = (1 + \tau\lambda)y_n$, верное при любом $n \geq 0$. Введём обозначение $\mu = \tau\lambda < 0$. Значение $\lambda < 0$ является параметром дифференциальной задачи (5.8), а τ есть параметр расчётной схемы Адамса. Условие монотонного стремления приближённого решения к нулю ($|y_{n+1}| < |y_n|$) будет выполняться, если $|1 + \mu| < 1$. Решив это неравенство, получаем условие $-2 < \mu < 0$.

Неравенство $-2 < \mu < 0$, так как $\mu = \tau\lambda < 0$, принимает вид $-2 < \tau\lambda$. Отсюда следует, что $\tau < -\dfrac{2}{\lambda} = \dfrac{2}{|\lambda|}$.

То есть, условие монотонного по n стремления к нулю приближенного решения, полученного по методу Адамса, будет выполнено, если τ — параметр расчётной схемы не превышает значения $\dfrac{2}{|\lambda|}$, определяемого параметром дифференциального уравнения.

Расчётные схемы, в которых величина шага дискретной сетки, параметр τ, связана с значениями параметров дифференциального уравнения, принято называть условно устойчивыми. Итак, явный одношаговый первого порядка аппроксимации метод Адамса является условно устойчивым.

Теперь применим для решения задачи (5.8) одношаговый метод Гира первого порядка аппроксимации (см. Пример 5.5.).

Расчётная схема метода имеет вид

$$\frac{y_{n+1} - y_n}{\tau} = f(t_{n+1}, y_{n+1}),$$

где правая часть $f(t_{n+1}, y_{n+1}) = \lambda y_{n+1}$.

Отсюда следует соотношение $y_{n+1}(1 - \tau\lambda) = y_n$, верное при любом $n \geq 0$. Используем обозначение $\mu = \tau\lambda < 0$. В данном случае, условие монотонного стремления приближённого решения к нулю ($|y_{n+1}| < |y_n|$) выполняется при любом $n \geq 0$, так как $|1 - \mu| > 1$.

То есть, при использовании одношагового метода Гира для решения модельной задачи (5.8) условие монотонности y_n выполнено при любых значениях τ и, следовательно, комбинация $\mu = \tau\lambda$ может быть равна любым значениям на отрицательной полуоси μ.

Если, при применении конкретной расчётной схемы для решения модельной задачи (5.8), условие монотонного по n стремления к нулю приближенного решения выполнено для любых значений параметра τ (то есть комбинация $\mu = \tau\lambda$ может принимать любые значения на отрицательной полуоси), то такую расчётную схему принято называть абсолютно устойчивой (A-устойчивой).

Итак, одношаговый метод Гира первого порядка аппроксимации является A-устойчивым или абсолютно устойчивым методом.

Проводить исследование m-шаговых методом при $m \geq 2$ на устойчивость удобно на модельной задаче (5.8), в которой параметр λ является произвольным комплексным числом. Тогда, комбинация $\mu = \tau\lambda$ также есть комплексное число и в этом случае используется следующее определение устойчивости расчётных схем.

Определение. Расчетная схема называется A(α)-устойчивой, если допустимые для этого метода значения μ принадлежат области $|arg(-\mu)| < \alpha$.

Следующие преобразования и рисунок иллюстрируют данное определение:

$$\mu = |\mu|\exp(i\arg\mu) \quad \Longrightarrow \quad -\mu = |\mu|\exp(i(\arg\mu - \pi)) \quad \Longrightarrow$$
$$|\arg(-\mu)| = |\arg\mu - \pi| < \alpha \quad \Longrightarrow \quad \pi - \alpha < \arg\mu < \pi + \alpha$$

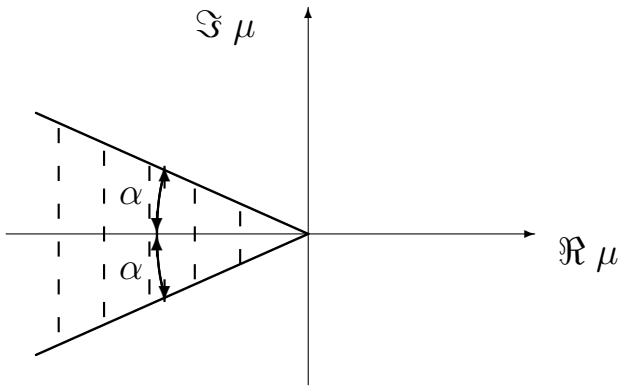

*Рис.*5.4

Верно утверждение (см. [3]):

Утверждение 5.1. Среди явных m-шаговых методов нет A(α)-устойчивых.

Среди неявных m-шаговых методов существуют A(α)-устойчивые методы. Например (см. [3]), четырехшаговый метод Гира четвертого порядка аппроксимации:

$$\frac{1}{12}(25y_n - 48y_{n-1} + 36y_{n-2} - 16y_{n-3} + 3y_{n-4}) = \tau f(t_n, y_n),\ n \geq 4.$$

5.5 Численное решение задачи Коши для системы обыкновенных дифференциальных уравнений

Рассмотрим задачу Коши для системы обыкновенных дифференциальных уравнений вида:

$$\begin{cases} \overline{u}_t = \overline{f}(t, \overline{u}), \quad t > 0; \\ \overline{u}\big|_{t=0} = \overline{u_0}, \end{cases}$$

где $\overline{u} = (u_1, u_2, \ldots, u_l)^T$ и $\overline{f} = (f_1, f_2, \ldots, f_l)^T$.

Расчетные схемы методов Рунге-Кутта (5.1) и m-шаговых методов (5.2) применимы и для численного решения систем обыкновенных дифференциальных уравнений. Продемонстрируем это на конкретном примере. Пусть задана задача Коши для двух обыкновенных дифференциальных уравнений:

$$\begin{cases} u_t = f(t, u, v); \\ v_t = g(t, u, v), \quad t > 0; \\ u(0) = u_0; \\ v(0) = v_0. \end{cases}$$

Применим для решения этой задачи численный метод Рунге-Кутта четвертого порядка аппроксимации (см. пункт 5.1). Введем обозначения $u(t_n) = y_n$ и $v(t_n) = w_n$. Тогда, расчетные формулы для вычисления y_n и w_n будут иметь вид:

$$\begin{cases} \dfrac{y_{n+1} - y_n}{\tau} = \dfrac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4); \\ \dfrac{w_{n+1} - w_n}{\tau} = \dfrac{1}{6}(M_1 + 2M_2 + 2M_3 + M_4), \end{cases}$$

а параметры $K_i,\ M_i,\ i = 1, 2, 3, 4$ вычисляются по следующим формулам:

$$\begin{array}{ll} K_1 = f(t_n, y_n, w_n), & M_1 = g(t_n, y_n, w_n); \\ K_2 = f(t_n + \frac{\tau}{2}, y_n + \frac{\tau}{2}K_1, w_n + \frac{\tau}{2}M_1), & M_2 = g(t_n + \frac{\tau}{2}, y_n + \frac{\tau}{2}K_1, w_n + \frac{\tau}{2}M_1); \\ K_3 = f(t_n + \frac{\tau}{2}, y_n + \frac{\tau}{2}K_2, w_n + \frac{\tau}{2}M_2), & M_3 = g(t_n + \frac{\tau}{2}, y_n + \frac{\tau}{2}K_2, w_n + \frac{\tau}{2}M_2); \\ K_4 = f(t_n + \tau, y_n + \tau K_3, w_n + \tau M_3), & M_4 = g(t_n + \tau, y_n + \tau K_3, w_n + \tau M_3). \end{array}$$

При численном решении систем обыкновенных дифференциальных уравнений проявляется существенное различие свойств абсолютно и условно устойчивых методов. Пусть дифференциальная задача Коши имеет вид:

$$\begin{cases} u_1' + a_1 u_1 = 0, \quad a_1 > 0; \\ u_2' + a_2 u_2 = 0, \quad a_2 > 0, \quad t > 0; \\ u_1(0) = u_2(0) = u_0. \end{cases}$$

Точными решениями этой задачи являются функции $u_1(t) = u_0 \exp(-a_1 t)$ и $u_2(t) = u_0 \exp(-a_2 t)$. Пусть выполняется неравенство $a_1 \gg a_2$ (в этом случае данная система является примером жесткой системы дифференциальных уравнений). Тогда, даже для $t = 1$ значение функции $u_1(1)$ будет много меньше значения функции $u_2(1)$ и близко к нулю. Если при численном решении этой задачи использовать условно устойчивый метод, например, явный одношаговый метод Адамса, то параметр расчётной схемы должен быть малой величиной, а именно $\tau < \frac{2}{a_1} \ll \frac{2}{a_2}$. Следовательно, для получения информации о значении функции $u_2(t)$ нужно будет выполнить большой объём «лишней» вычислительной работы. Поэтому, при численном решении жестких систем обыкновенных дифференциальных уравнений рекомендуется пользоваться A(α)-устойчивыми расчётными схемами.

Вспомним определение жесткой системы обыкновенных дифференциальных уравнений. Пусть имеется задача:

$$\begin{cases} \overline{u}_t &= \overline{f}(t, \overline{u}), \quad t > 0; \\ \overline{u}\big|_{t=0} &= \overline{u_0}, \end{cases}$$

Поставим в соответствие этой системе обыкновенных дифференциальных уравнений матрицу, элементами которой являются частные производные всех правых частей уравнений по всем неизвестным функциям (якобиан)

$$A(t, \overline{u}) = \left(\frac{\partial f_j(t, \overline{u})}{\partial u_i} \right), \quad i,\ j = 1,\, 2,\, \ldots,\, l.$$

Пусть $\lambda_k(t),\ t \in [0;\, T]$ — собственные значения этой матрицы.

Определение. Система дифференциальных уравнений называется жесткой, если выполнены условия:

$$\operatorname{Re} \lambda_k(t) < 0 \quad \forall\, k,\ t \in [0;\, T];$$

$$\sup_{t \in [0;\, T]} \frac{\max\limits_k |\operatorname{Re} \lambda_k(t)|}{\min\limits_k |\operatorname{Re} \lambda_k(t)|} \gg 1.$$

Глава 6

Численное решение краевой задачи для обыкновенного дифференциального уравнения

Рассмотрим специфику численного решения краевых задач для обыкновенных диффференциальных уравнений на примере следующей краевой задачи:

$$\begin{cases} (k(x)u'(x))' - q(x)u(x) + f(x) = 0, \quad 0 < x < l; \\ -k(0)u'(0) + \beta u(0) = \mu_1; \\ u(l) = \mu_2. \end{cases} \tag{6.1}$$

Здесь $k(x),\ q(x),\ f(x)$ – заданные гладкие функции, для которых выполняются условия $k(x) \geqslant k_0 > 0,\ q(x) \geqslant 0,\ \beta \geqslant 0,$ а $\mu_1,\ \mu_2$ – заданные числа. При таких условиях решение задачи (6.1) существует и единственно.

Действия для получение численного решения задачи (6.1) можно разделить на несколько этапов. Сначала дифференциальной задаче сопоставляется её алгебраический аналог (разностная схема).

6.1 Разностная схема

Существуют различные методы построения разностных схем. Первым рассмотрим интегро-интерполяционный метод.

6.1.1 Интегро-интерполяционный метод построения разностной схемы

Название метода связано стем, что при построении разностной схемы осуществляется переход от дифференциального уравнения к его интегральному аналогу, который заменяется приближёнными квадратурными формулами.

Прежде всего на отрезке $x \in [0;\, l]$ введем равномерную дискретную сетку

$$\omega_h = \{x_i = ih,\ i = \overline{0,\, N},\ h = \frac{l}{N}\}.$$

Средние точки между узлами дискретной сетки обозначим как $x_{i\pm\frac{1}{2}} = x_i \pm \dfrac{h}{2}$. Пусть $W(x) = k(x)u'(x)$.

Фиксируем произвольное $i \in [1;\, N-1]$ и интегрируем дифференциальное уравнение в задаче (6.1) на отрезке $x \in [x_{i-\frac{1}{2}};\, x_{i+\frac{1}{2}}]$:

$$\int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} ((k(x)u'(x))' - q(x)u(x) + f(x))\,dx = 0 \iff$$

$$W_{i+\frac{1}{2}} - W_{i-\frac{1}{2}} - \int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q(x)u(x)\,dx + \int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x)\,dx = 0, \tag{6.2}$$

где $W_{i\pm\frac{1}{2}} = W(x_{i\pm\frac{1}{2}})$.

Используя теорему о среднем значении, интеграл $\displaystyle\int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q(x)u(x)\,dx$ можно приближенно заменить выражением $u_i \displaystyle\int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q(x)\,dx$, где $u_i = u(x_i)$. Введем обозначения $\varphi_i = \dfrac{1}{h}\displaystyle\int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x)\,dx$ и $d_i = \dfrac{1}{h}\displaystyle\int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q(x)\,dx$. Тогда, (6.2) переходит в приближенное равенство

$$W_{i+\frac{1}{2}} - W_{i-\frac{1}{2}} - hd_iu_i + h\varphi_i \approx 0. \tag{6.3}$$

Так как $u'(x) = \dfrac{W(x)}{k(x)}$, то интегрируя это равенство на отрезке $x \in [x_i;\, x_{i+1}]$, получим $u_{i+1} - u_i = \displaystyle\int_{x_i}^{x_{i+1}} \frac{W(x)}{k(x)}\,dx$. Используя теорему о среднем значении, заменим равенство приближенным соотношением $u_{i+1} - u_i \approx W_{i+\frac{1}{2}} \displaystyle\int_{x_i}^{x_{i+1}} \frac{dx}{k(x)}$. Введём обозначение $a_{i+1} = \left[\dfrac{1}{h}\displaystyle\int_{x_i}^{x_{i+1}} \frac{dx}{k(x)}\right]^{-1}$.

Тогда приближенное равенство примет вид $W_{i+\frac{1}{2}} \approx a_{i+1}\dfrac{u_{i+1}-u_i}{h}$. Подставляя приближенное выражение для $W_{i+\frac{1}{2}}$ и $W_{i-\frac{1}{2}}$ в (6.3), получим

$$a_{i+1}\frac{u_{i+1}-u_i}{h} - a_i\frac{u_i-u_{i-1}}{h} - hd_iu_i + h\varphi_i \approx 0. \tag{6.4}$$

Обозначим через y_i значения сеточной функции, которые превращают (6.4) в точное равенство

$$a_{i+1}\frac{y_{i+1}-y_i}{h} - a_i\frac{y_i-y_{i-1}}{h} - hd_iy_i + h\varphi_i = 0. \tag{6.5}$$

Эти y_i будем считать приближенными значениями искомой функции $u(x)$ в точках x_i.

Запишем (6.5) в компактном стандартном виде. Используя обозначения $y_{\overline{x},i+1} = \dfrac{y_{i+1}-y_i}{h}$ и $y_{\overline{x},i} = \dfrac{y_i-y_{i-1}}{h}$, приводим (6.5) к виду $a_{i+1}y_{\overline{x},i+1} - a_iy_{\overline{x},i} - hd_iy_i + h\varphi_i = 0$ или $(ay_{\overline{x}})_{i+1} - (ay_{\overline{x}})_i - hd_iy_i + h\varphi_i = 0$. Разделённая разность $\dfrac{(ay_{\overline{x}})_{i+1}-(ay_{\overline{x}})_i}{h} = (ay_{\overline{x}})_{x,i}$ — является разностной производной вперед комбинации $(ay_{\overline{x}})$ в точке x_i.

Тогда (6.5) принимает вид

$$(ay_{\overline{x}})_{x,i} - d_iy_i + \varphi_i = 0. \tag{6.6}$$

Уравнения (6.6), записанные для всех $i = \overline{1,\, N-1}$, представляют собой алгебраический аналог дифференциального уравнения в задаче (6.1) и являются системой линейных алгебраических уравнений относительно

$y_i, \quad i = \overline{0,\, N}$.

Эта система состоит из $(N-1)$ уравнений относительно $(N+1)$ неизвестных. Два дополнительных уравнения, необходимых для формального замыкания системы линейных алгебраических уравнений, получим используя краевые условия в задаче (6.1).

Краевое условие $u(l) = \mu_2$ дает одно дополнительное уравнение

$$y_N = \mu_2. \tag{6.7}$$

Для получения второго дополнительного алгебраического уравнения проинтегрируем дифференциальное уравнение задачи (6.1) по отрезку $x \in [0; \frac{h}{2}]$.

В результате получим

$$W_{\frac{1}{2}} - k(0)u'(0) - u_0 \int\limits_0^{\frac{h}{2}} q(x)\,dx + \int\limits_0^{\frac{h}{2}} f(x)\,dx \approx 0. \tag{6.8}$$

Введём обозначения $\varphi_0 = \dfrac{1}{\frac{h}{2}} \displaystyle\int\limits_0^{\frac{h}{2}} f(x)\,dx$, $d_0 = \dfrac{1}{\frac{h}{2}} \displaystyle\int\limits_0^{\frac{h}{2}} q(x)\,dx$. Так как $W_{\frac{1}{2}} \approx a_1 \dfrac{u_1 - u_0}{h}$ и $k(0)u'(0) = \beta u_0 - \mu_1$, то, требуя точного выполнения равенства в (6.8), получим

$$a_1 y_{x,0} - \beta y_0 + \mu_1 - \frac{h}{2} d_0 y_0 + \frac{h}{2}\varphi_0 = 0$$

или

$$-a_1 y_{x,0} + \overline{\beta} y_0 = \overline{\mu}_1, \tag{6.9}$$

где $\overline{\beta} = \beta + \dfrac{h}{2} d_0$, $\overline{\mu}_1 = \mu_1 + \dfrac{h}{2}\varphi_0$.

Совокупность уравнений (6.6), (6.7) и (6.9) представляет собой разностную схему, соответствующую исходной краевой задаче (6.1) для обыкновенного дифференциального уравнения. Данная разностная схема является системой линейных алгебраических уравнений, которую можно компактно записать в виде

$$Ry = g, \tag{6.10}$$

где R — квадратная матрица, в которой $(N+1)$ столбцов и $(N+1)$ строк,

$y = (y_0, y_1, \ldots, y_i, \ldots, y_{N-1}, y_N)^T$ — вектор неизвестных

и $g = (\overline{\mu}_1, \varphi_1, \ldots, \varphi_i, \ldots, \varphi_{N-1}, \mu_2)^T$ — вектор заданных правых частей.

6.1.2 Метод аппроксимации квадратичного функционала

Рассмотрим еще один метод построения разностных схем. Пусть в краевой задаче (6.1) заданы параметры $l = 1$, $\beta = 1$, $\mu_1 = \mu_2 = k(0) = 0$. Введём обозначение для дифференциального оператора $Lu = -(k(x)u'(x))' + q(x)u(x)$.

Тогда (6.1) принимает вид

$$\begin{cases} Lu(x) = f(x), \quad 0 < x < 1; \\ u(0) = u(1) = 0. \end{cases} \tag{6.11}$$

Пусть для функций $u(x)$ и $v(x)$, заданных на отрезке $x \in [0;\, 1]$, скалярное произведение $(u, v) = \int\limits_0^1 u(x)v(x)\,dx$. Тогда, скалярное произведение $(Lu, u) = \int\limits_0^1 \left(k(x)(u'(x))^2 + q(x)u^2(x)\right)\,dx > 0$ при $u(0) = u(1) = 0$ для любой функции $u(x)$ не равной тождественно нулю на отрезке $x \in [0;\, 1]$.

В силу положительности дифференциального оператора L, решение краевой задачи (6.11) эквивалентно поиску функции $u(x)$, минимизирующей функционал

$$J\,(u) = (Lu, u) - 2\,(f, u) = \int\limits_0^1 \left(k(x)(u'(x))^2 + q(x)u^2(x) - 2f(x)u(x)\right)\,dx.$$

Введём на отрезке $x \in [0;\, 1]$ равномерную дискретную сетку $\omega_h = \{x_i = ih,\ i = \overline{0,\, N},\ h = \dfrac{1}{N}\}$. Тогда

$$J\,(u) = \sum_{i=1}^{N} \int\limits_{x_{i-1}}^{x_i} \left(k(x)(u'(x))^2 + q(x)u^2(x) - 2f(x)u(x)\right)\,dx.$$

На каждом отрезке $x \in [x_{i-1};\, x_i]$ производную $u'(x)$ приближённо заменим константой $u_{\overline{x},i}$, равной первой разностной производной назад в точке x_i. Введем обозначение $a_i = \dfrac{1}{h}\int\limits_{x_{i-1}}^{x_i} k(x)\,dx$. При этом

$$J\,(u) \approx \sum_{i=1}^{N} \left(a_i u_{\overline{x},i}^2 h + \int\limits_{x_{i-1}}^{x_i} \left(q(x)u^2(x) - 2f(x)u(x)\right)\,dx \right).$$

Вычислим оставшиеся интегралы приближённо, используя квадратурную формулу трапеций. В результате получим

$$J\,(u) \approx J_h(u_0 = 0,\, u_1,\, \ldots,\, u_{N-1},\, u_N = 0) =$$

$$= \sum_{i=1}^{N} \left(a_i u_{\overline{x},i}^2 h + \frac{1}{2}(q_i u_i^2 - 2 f_i u_i + q_{i-1} u_{i-1}^2 - 2 f_{i-1} u_{i-1}) h \right) =$$
$$= \sum_{i=1}^{N} a_i u_{\overline{x},i}^2 h + \sum_{i=1}^{N-1} (q_i u_i^2 - 2 f_i u_i) h.$$

Итак, минимизация функционала $J(u)$ сведена к поиску минимума J_h — функции многих переменных. Необходимым условием минимума является равенство нулю частных производных $\dfrac{\partial J_h}{\partial u_i}$ для всех $i = \overline{1,\, N-1}$. Вычислим производные:

$$\frac{\partial J_h}{\partial u_i} = 2 a_{i+1} u_{\overline{x},i+1} (-\frac{1}{h}) h + 2 a_i u_{\overline{x},i} (\frac{1}{h}) h + (2 q_i u_i - 2 f_i) h =$$
$$= -2h \left(\frac{a_{i+1} u_{\overline{x},i+1} - a_i u_{\overline{x},i}}{h} - q_i u_i + f_i \right) = -2h \left((a u_{\overline{x}})_{x,i} - q_i u_i + f_i \right).$$

Приравнивая эти выражения к нулю, получим систему алгебраических уравнений относительно y_i $(i = \overline{0,\, N})$ следующего вида:

$$\begin{cases} (a y_{\overline{x}})_{x,i} - q_i y_i + f_i = 0, \; i = \overline{1,\, N-1} \\ y_0 = y_N = 0. \end{cases}$$

Данная система линейных алгебраических уравнений представляет собой разностную схему, соответствующую краевой задаче (6.11).

6.2 Элементы теории разностных схем

После построения для краевой задачи (6.1) разностной схемы (6.6), (6.7), (6.9) появляются вопросы о существовании, единственности решения разностной схемы, выборе метода отыскания этого решения и связи между решением разностной схемы с точным решением краевой задачи (6.1).

6.2.1 Решение разностной схемы

Рассматриваемая разностная схема представляет собой систему линейных алгебраических уравнений, которую можно записать в виде

$$A_i y_{i-1} - C_i y_i + B_i y_{i+1} = -F_i, \quad i = \overline{1,\, N-1},$$
$$y_0 = \kappa_1 y_1 + \nu_1, \quad y_N = \mu_2,$$

где

$$A_i = a_i,\ B_i = a_{i+1},\ C_i = a_i + a_{i+1} + d_i h^2,\ F_i = \varphi_i h^2,$$
$$\kappa_1 = \frac{a_1}{a_1 + \overline{\beta} h},\ \nu_1 = \frac{\overline{\mu}_1}{a_1 + \overline{\beta} h}.$$

Оптимальным методом решения таких систем линейных алгебраических уравнений является метод прогонки.

Для коэффициентов уравнений выполнены условия $|C_i| \geq |A_i| + |B_i|$, $A_i \neq 0$, $B_i \neq 0$ $(i = \overline{1,\ N-1})$, которые гарантируют существование, единственность решения разностной схемы и возможность применения и устойчивость метода прогонки (см. [3]).

6.2.2 Порядок аппроксимации

Разностная схема представляет собой систему алгебраических уравнений, каждое из которых сопоставляется дифференциальной задачи в конкретном узле разностной сетки. Величина шага h (расстояние между соседними узлами) разностной сетки фактически определяет размерность алгебраической системы. Чем меньше h, тем больше узлов разностной сетки и тем выше размерность алгебраической системы.

Решение разностной схемы (6.10), сеточная функция $y_i\,(i = \overline{0,\ N})$, считается приближённым значением для $u_i = u(x_i)\,(i = \overline{0,\ N})$ — точного решения дифференциальной задачи (6.1). Сеточная функция y_i может не совпадать с значением точного решения u_i в узле x_i. Поэтому, формальная подстановка в разностную схему значений u_i вместо y_i может нарушать точные равенства в разностной схеме. Возникающий дисбаланс в каждом алгебраическом уравнении называется погрешностью аппроксимации (невязкой) разностной схемой дифференциальной задачи, вычисленной на точном решении дифференциальной задачи. Погрешность аппроксимации в узле x_i будем обозначать через ψ_i. Всю совокупность значений сеточной функции $\psi_i\,(i = \overline{0,\ N})$ удобно характеризовать одной количественной величиной, называемой нормой и обозначаемой $\|\psi\|$. За норму сеточной функции принимается неотрицательное число, определяющее величину отклонения сеточной функции от тождественного нуля. Примером нормы является величина

$$||\psi||_h = \max_i |\psi_i|.$$

Определение. Разностная схема аппроксимирует дифференциальную задачу на всей дискретной сетке, если норма погрешности аппроксимации стремится к нулю с уменьшением h :

$$||\psi||_h = \max_i |\psi_i| \xrightarrow{h\to 0} 0.$$

Если, кроме того, выполняется неравенство $||\psi||_h \leq ch^p \quad (||\psi||_h = O(h^p))$, где $c > 0$ и $p > 0$ — некоторые постоянные не зависящие от h, то используется следующий термин.

Определение. Разностная схема аппроксимирует дифференциальную задачу с p-м порядком аппроксимации.

Можно показать (см. [3]), что разностная схема, построенная с помощью интегро-интерполяционного метода, аппроксимирует дифференциальную задачу (6.1) с вторым порядком аппроксимации. Используя этот факт, допустимо при вычислении интегралов, определяющих a_i, φ_i и d_i, использовать приближённые квадратурные формулы второго порядка точности.

Так, например, используя квадратурную формулу прямоугольников, получим для параметров разностной схемы следующие формулы $a_i = k\left(x_{i-\frac{1}{2}}\right)$, $d_i = q(x_i)$ и $\varphi_i = f(x_i)$. Применяя квадратурную формулу трапеций, получим соотношения $a_i = \dfrac{2k(x_i)k(x_{i-1})}{k(x_{i-1}) + k(x_i)}$, $d_i = \dfrac{1}{2}\left(q\left(x_{i-\frac{1}{2}}\right) + q\left(x_{i+\frac{1}{2}}\right)\right)$ и $\varphi_i = \dfrac{1}{2}\left(f\left(x_{i-\frac{1}{2}}\right) + f\left(x_{i+\frac{1}{2}}\right)\right)$.

6.2.3 Устойчивость разностной схемы

Рассмотрим, используемое в теории разностных схем, понятие устойчивости разностной схемы. Пусть сеточная функция $y_i\,(i = \overline{0,\, N})$ является решением разностной схемы (6.10). Это решение соответствует заданным значениям сеточной функции $g_i\,(i = \overline{0,\, N})$. Внесем возмущения $\delta g_i\,(i = \overline{0,\, N})$ в значения функции $g_i\,(i = \overline{0,\, N})$ и решим разностную схему с возмущённой сеточной функцией $g_i + \delta g_i$, $(i = \overline{0,\, N})$. Функция $y_i + \delta y_i\,(i = \overline{0,\, N})$, являющаяся решением возмущённой разностной схемы, может отличаться от функции $y_i\,(i = \overline{0,\, N})$ — решения невозмущённой разностной схемы.

Определение. Разностная схема устойчива, если выполнено неравенство

$$||\delta y|| \leqslant c||\delta g||, \tag{6.12}$$

где $c > 0$ — некоторая константа, не зависящая от параметров разностной схемы.

Неравенство (6.12) означает, что малое возмущение параметров разностной схемы вызывает малое возмущение решения разностной схемы. Данное определение устойчивости используется как для линейных, так и для нелинейных разностных схем.

Если разностная схема является линейной системой алгебраических уравнений, то предыдущее определение устойчивости эквивалентно следующему определению.

Определение. Линейная разностная схема (6.10) устойчива, если выполнено неравенство

$$||y|| \leqslant c||g||, \tag{6.13}$$

где $c > 0$ — некоторая константа, не зависящая от параметров разностной схемы.

Покажем равносильность обоих определений устойчивости для линейной разностной схемы (6.10) ([4]).

Пусть линейная разностная схема $Ry = g$ имеет единственное решение, которое удовлетворяет неравенству $||y|| \leqslant c||g||$ (6.13). Обозначим через y_1 решение разностной схемы $Ry_1 = g + \delta g$ с возмущённой правой частью. Разность $\delta y = y_1 - y$, в силу линейности разностной схемы, является решением системы $R\delta y = \delta g$ и для этого решения, в силу (6.13), выполняется неравенство $||\delta y|| \leqslant c||\delta g||$ (6.12).

Покажем обратное. Пусть δy — возмущение решения линейной разностной схемы $Ry = g$. Это возмущение является откликом на внесение произвольного возмущения δg в правую часть разностной схемы, то есть $R(y + \delta y) = g + \delta g$. Пусть для δy и δg выполнено неравенство $||\delta y|| \leqslant c||\delta g||$ (6.12). Тогда, в силу линейности разностной схемы, δy является решением разностной схемы $R\delta y = \delta g$ для любого δg. Пусть $\delta g = g$. Тогда $\delta y = y$ и, в силу (6.12), верно неравенство $||y|| \leqslant c||g||$ (6.13).

Итак, для линейной разностной схемы, оба определения устойчивости разностной схемы эквивалентны.

Можно показать (см. [3]), что линейная разностная схема (6.10), построенная с помощью интегро-интерполяционного метода, является устойчивой разностной схемой, так как для её решения выполняется неравенство $||y||_h \leqslant c||g||_h$.

6.2.4 Сходимость решения разностной схемы

Решение разностной схемы $Ry = g$ (6.10), сеточная функция y_i $(i = \overline{0,\ N})$, при уменьшении шага дискретной сетки $(h \longrightarrow 0\ (N \longrightarrow \infty))$, представляет собой всё более подробную таблицу числовых значений y_i. Необходимо уметь сопостовлять эти числовые данные с точным решением дифференциальной задачи (6.1).

Сеточную функцию $z_i = y_i - u_i$ $(i = \overline{0,\ N})$ назовём погрешностью решения разностной схемы.

Определение. Приближенное решение y_i сходится к точному решению дифференциальной задачи на всей дискретной сетке, если норма погрешности решения разностной схемы стремится к нулю с уменьшением h :

$$||z||_h = \max_i |z_i| \xrightarrow{h\to 0} 0.$$

Если, кроме того, выполняется неравенство $||z||_h \le ch^p \quad (||z||_h = O(h^p))$, где $c > 0$ и $p > 0$ — некоторые постоянные не зависящие от h, то используется следующий термин.

Определение. Решение разностной схемы имеет p-ый порядок точности (сходится с p-м порядком.)

Исследуем решение разностной схемы (6.10) на сходимомть. Подставим в разностную схему $Ry = g$ (6.10) её решение, записанное в виде $y_i = u_i + z_i\, (i = \overline{0,\, N})$. Тогда получим, $Rz = -Ry + g$ или $Rz = \psi$, где ψ — погрешность аппроксимации. Так как разностная схема (6.10) устойчива (см. [3]), то верно неравенство $||z||_h \le c||\psi||_h$.

Разностная схема (6.10) аппроксимирует дифференциальную задачу (6.1) с 2-м порядком по h (см. [3]), то есть $||\psi||_h \le c_1 h^2$.

Тогда верно неравенство $||z||_h \le cc_1 h^2$.

Следовательно, решение разностной схемы (6.10) сходится к точному решению задачи (6.1) и имеет 2-ой порядок точности по h.

Для любой линейной разностной схемы справедливо следующее утверждение.

Теорема 6.1. Пусть линейная разностная схема аппроксимирует дифференциальную задачу и устойчива. Тогда решение разностной схемы сходится к точному решению дифференциальной задачи и порядок точности решения разностной схемы совпадает с порядком аппроксимации.

Доказательство. Так как разностная схема устойчива, то верно неравенство $||y|| \le c||g||$, где y — решение разностной схемы и g — вектор правых частей в линейной разностной схеме. В силу линейности разностной схемы, z — погрешность решения разностной схемы является решением той же разностной схемы с правой частью равной ψ (погрешность аппроксимации). Тогда, в силу устойчивости разностной схемы, выполняется неравенство $||z|| \le c||\psi||$.

Так как разностная схема аппроксимирует дифференциальную задачу с p-ым порядком, то $||\psi|| \le c_1 h^p$.

Тогда верно неравенство $||z|| \le cc_1 h^p$.

Отсюда следует, что решение разностной схемы сходится к точному решению дифференциальной задачи и порядок точности решения разностной схемы совпадает с порядком аппроксимации. □

При исследовании свойств любой линейной разностной схемы, во-первых, устанавливают наличие аппроксимации. Во-вторых, показывают устойчивость разностной схемы, а затем, используя доказанную теорему, делают вывод о наличие сходимости решения разностной схемы к точному решению дифференциальной задачи и устанавливают порядок точности решения разностной схемы.

Глава 7

Численное решение краевых задач для дифференциальных уравнений в частных производных

В этой главе обсуждаются численные методы решения краевых задач для дифференциальных уравнений в частных производных второго порядки. В основе всех этих численных методов лежит сопоставление дифференциальным уравнениям систем линейных алгебраических уравнений. В соответствии с принятой в уравнениях математической физики классификацией дифференциальных уравнений, рассматриваются методы численного решения краевых задач для уравнений параболического, гиперболического и эллиптического типа.

7.1 Разностные схемы для уравнений параболического типа

Для устранения лишних технических сложностей, рассмотрим численное решение простейших краевых задач, для которых решение можно получить и известными аналитическими методами. Рассматриваемая техника численного решения краевых задач легко обобщается на случаи, для которых неприменимы существующие аналитические методы решения.

7.1.1 Явная разностная схема для уравнения теплопроводности

В качестве примера краевой задачи для уравнения параболического типа, рассмотрим первую краевую задачу для одномерного по пространству уравнения теплопроводности.

Пусть функция $u(x,t)$ является решением следующей краевой задачи:

$$\begin{cases} u_t(x,t) = u_{xx}(x,t) + f(x,t), & 0 < x < l,\ 0 < t; \\ u(x,0) = v(x), & 0 \leqslant x \leqslant l; \\ u(0,t) = \mu_1(t), & 0 < t; \\ u(1,t) = \mu_2(t), & 0 < t. \end{cases} \tag{7.1}$$

Для построения численного решения задачи (7.1) введём ограничение на переменную t. Пусть $0 \leq t \leq T$, где T заданное число. Проведем дискретизацию области изменения независимых переменных x и t. На отрезках $x \in [0;\, l]$ и $t \in [0;\, T]$ введём дискретные наборы точек $\omega_h = \left\{x_j = jh,\ j = \overline{0,N},\ h = \frac{l}{N}\right\}$ и $\omega_\tau = \left\{t_n = n\tau,\ n = \overline{0,K},\ \tau = \frac{T}{K}\right\}$. Значения параметров N и K характеризуют «подробность» ноборов точек. В плоскости переменных $(x,\, t)$ проведём линии $x = x_j$ $(j = \overline{0,N})$ и $t = t_n$ $(n = \overline{0,K})$ (см. Рис.7.1). Точки пересечения этих линий будем называть узлами дискретной сетки и поделим их на две группы. Внутренние узлы (точки помеченные кругами), это точки в которых выполняется дифференциальное уравнение задачи (7.1). Граничные узлы (точки помеченные крестами), это точки в которых выполняются начальное или граничные условия задачи (7.1).

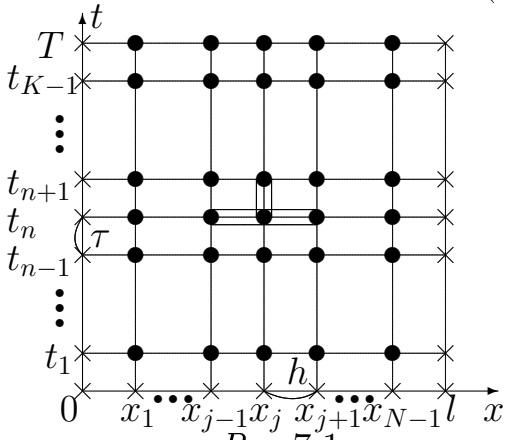

Рис.7.1

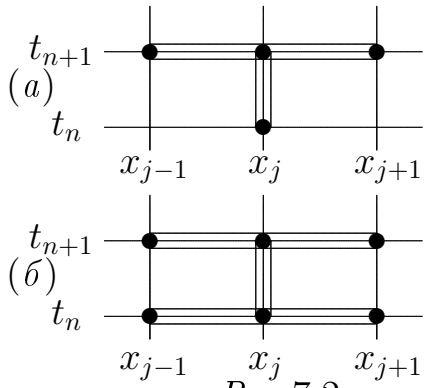

Рис.7.2

Дискретные узлы с одинаковыми значениями координаты t_n будем называть временным слоем.

Введём обозначения $u(x_j,\, t_n) = u_j^n$, $f(x_j,\, t_n) = f_j^n$, $v(x_j) = v_j$, $\mu_1(t_n) = \mu_1^n$ и $\mu_2(t_n) = \mu_2^n$. В дифференциальном уравнении задачи (7.1) присутствуют производные $u_t(x,t)$ и $u_{xx}(x,t)$, числовые значения которые во внутренних узлах дискретной сетки приближенно равны значениям алгебраических выра-

жений $u_t(x_j,\, t_n) \approx \dfrac{u_j^{n+1} - u_j^n}{\tau}$ и $u_{xx}(x_j,\, t_n) \approx \dfrac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2}$. Комбинации, приближающие производные, используют значения функции $u(x,\, t)$ в трёх узлах дискретной сетки на $n-$ом временном слое и в одном узле на $(n+1)-$ом временном слое (см. Рис.7.1).

Определение. Совокупность узлов дискретной сетки, на базе которых построено алгебраическое приближение дифференциального уравнения, называется шаблоном разностной схемы.

В каждом внутреннем узле дискретной сетки запишем комбинацию алгебраических выражений для производных, повторяющую структуру дифференциального уравнения, и потребуем точного выполнения этих алгебраических соотношений. В граничных узлах запишем соответствующие начальные и граничные условия. В результате получим следующую систему линейных алгебраических уравнений:

$$\begin{cases} \dfrac{y_j^{n+1} - y_j^n}{\tau} = \dfrac{y_{j+1}^n - 2y_j^n + y_{j-1}^n}{h^2} + \varphi_j^n,\ j = \overline{1,\, N-1},\ n = \overline{0,\, K-1}; \\ y_j^0 = v_j, \quad j = \overline{0,\, N}; \\ y_0^n = \mu_1^n, \quad n = \overline{1, K}; \\ y_N^n = \mu_2^n, \quad n = \overline{1, K}. \end{cases} \tag{7.2}$$

Здесь y_j^n $(j = \overline{0,\, N},\ n = \overline{0,\, K})$ искомые приближенные значения функции $u(x,\, t)$, а φ_j^n заданные приближённые значения функции $f(x,\, t)$ во внутренних узлах дискретной сетки. Точность вычисления φ_j^n обсудим позже.

Система (7.2) представляет собой разностную схему для задачи (7.1). Эта разностная схема построена с использованием 4-ёх точечного шаблона, состоящего из трёх узлов дискретной сетки на $n-$ом временном слое и одного узла на $(n+1)-$ом временном слое (см. Рис.7.1).

Система (7.2) состоит из $((N-1)K + (N+1) + 2K = NK + N + K + 1)$ уравнений. Неизвестными в этой алгебраической системе являются y_j^n $(j = \overline{0,\, N},\ n = \overline{0,\, K})$. Общее число неизвестных равно $((N+1)(K+1) = NK + N + K + 1)$.

Введём в рассмотрение вектор y, компонентами которого являются неизвестные y_j^n, упорядоченные по возрастанию номера временного слоя и при одинаковом n по возрастающему индексу j, то есть $y = (y_0^0,\, y_1^0,\, \ldots, y_N^0,\, y_0^1,\, y_1^1,\, \ldots, y_N^1,\, \ldots\ \ldots\, y_0^{K-1},\, y_1^{K-1},\, \ldots, y_N^{K-1},\, y_0^K,\, y_1^K,\, \ldots, y_N^K)^T$. Тогда система уравнений (7.2) в матричной форме записи примет вид

$$Ry = g, \tag{7.3}$$

где вектор правых частей равен $g = (v_0, v_1, \ldots, v_N,\ \mu_1^1, \varphi_1^0, \varphi_2^0, \ldots, \varphi_{N-1}^0, \mu_2^1,\ \mu_1^2, \varphi_1^1, \varphi_2^1, \ldots, \varphi_{N-1}^1, \mu_2^2, \ldots\ \ \ldots \mu_1^K, \varphi_1^{K-1}, \varphi_2^{K-1}, \ldots, \varphi_{N-1}^{K-1}, \mu_2^K)^T$, а квадратная матрица R большой размерности имеет «ленточный» вид (см. Рис.7.3), соответствующий выбранному порядку расположения неизвестных в компонентах вектора y.

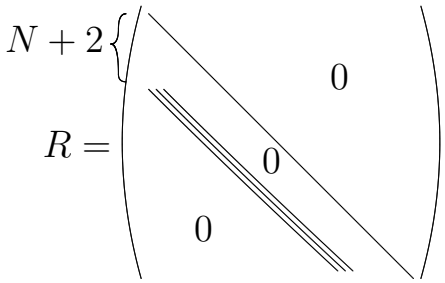

*Рис.*7.3

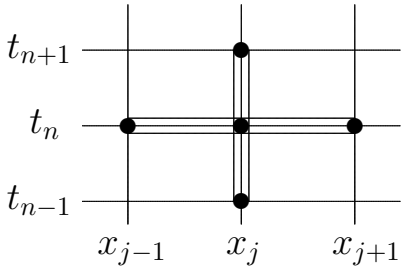

*Рис.*7.4

На главной диагонали матрицы R находятся числа $\left(\frac{1}{\tau}\right)$ или единица. На остальных трех диагоналях расположены значения комбинаций $\left(-\frac{1}{h^2}\right)$, $\left(-\frac{1}{\tau}+\frac{2}{h^2}\right)$ и ряд нулей. Вне этих четырех диагоналей все элементы матрицы R равны нулю.

Так как матрица R является нижней треугольной матрицей, то решение системы (7.2) проводится по явным формулам пересчета, начиная с первого уравнения. Поэтому, разностную схему (7.2) называют явной разностной схемой.

При использовании любой разностной схемы необходимо получить ответы на следующие вопросы:

1) есть ли аппроксимация разностной схемой исходной дифференциальной задачи и каков порядок аппроксимации;

2) существует ли единственное решение разностной схемы и как его найти;

3) является ли разностная схема устойчивой.

Получив утвердительные ответы на эти вопросы, можно делать выводы о сходимости решения разностной схемы к точному решению дифференциальной задачи и о порядке точности полученного решения разностной схемы.

Исследуем аппроксимацию, разрешимость и устойчивость разностной схемы (7.2).

Порядок аппроксимации.

В граничных узлах дискретной сетки начальное и граничные условия задачи выполняются точно. Во внутренних узлах сетки невязка в разностной схеме (7.2) имеет вид

$$\psi_j^n = -\frac{u_j^{n+1}-u_j^n}{\tau}+\frac{u_{j+1}^n-2u_j^n+u_{j-1}^n}{h^2}+\varphi_j^n,\ j=\overline{1,\,N-1},\ n=\overline{0,K-1}. \quad (7.4)$$

Представим $u_{j\pm1}^n$ и u_j^{n+1} в виде разложений в ряд Тейлора

$$\begin{aligned} u_{j\pm1}^n &= u_j^n \pm u_{x,j}^n h + u_{xx,j}^n\frac{h^2}{2} \pm u_{xxx,j}^n\frac{h^3}{6}+O(h^4);\\ u_j^{n+1} &= u_j^n + u_{t,j}^n\tau + O(\tau^2) \end{aligned}$$

и подставим в (7.4). Тогда выражение для невязки примет вид

$$\psi_j^n = (-u_{t,j}^n + u_{xx,j}^n)+\varphi_j^n + O(\tau+h^2).$$

Прибавим и вычтем в правой части значение функции f_j^n. В результате

$$\psi_j^n = (-u_{t,j}^n + u_{xx,j}^n + f_j^n)+\varphi_j^n - f_j^n + O(\tau+h^2).$$

Так как значение u_j^n является точным решением уравнения теплопроводности (7.1), то комбинация в скобках в правой части равна нулю. Если для правой части разностной схемы (7.2) выполнено соотношение $\varphi_j^n = f_j^n+O(\tau+h^2)$, то невязка будет равна $\psi_j^n = O(\tau+h^2),\ j=\overline{1,\,N-1},\ n=\overline{0,K-1}$. Отсюда следует, что

$$||\psi||_{\tau,h} = \max_{j,n}|\psi_j^n| = O(\tau+h^2).$$

Итак, разностная схема (7.2) аппроксимирует краевую задачу (7.1) с первым порядком по τ и вторым порядком аппроксимации по h.

Метод решения разностной схемы.

Преобразуем разностное уравнение (7.2) к виду

$$y_j^{n+1} = y_j^n + \frac{\tau}{h^2}\left(y_{j+1}^n - 2y_j^n + y_{j-1}^n\right)+\tau\varphi_j^n,\ j=\overline{1,\,N-1},\ n=\overline{0,K-1} \quad (7.5)$$

и будем считать его уравнением относительно y_j^{n+1}.

Выбрав в (7.5) значение $n=0$, получим

$$y_j^1 = v_j + \frac{\tau}{h^2}\left(v_{j+1}-2v_j+v_{j-1}\right)+\tau\varphi_j^0,\quad j=\overline{1,\,N-1}.$$

По этой формуле для всех $j = \overline{1,\, N-1}$ вычисляются на первом временном слое $t = \tau$ искомые значения y_j^1.

Пусть в (7.5) значение $n = 1$. Тогда

$$y_j^2 = y_j^1 + \frac{\tau}{h^2}\left(y_{j+1}^1 - 2y_j^1 + y_{j-1}^1\right) + \tau\varphi_j^1,\ j = \overline{1,\, N-1}.$$

В правой части равенства значения y_j^1 уже найдены. Поэтому можем вычислить искомые y_j^2 на всем втором временном слое $t = 2\tau$. Аналогично можно вычислить y_j^n на третьем и всех последующих временных слоях.

Таким образом, данная разностная схема решается по временным слоям и решение разностной схемы (7.2) существует и единственно.

Устойчивость разностной схемы.

Рассмотрим вопрос об устойчивости разностной схемы (7.2). Используем матричную форму записи разностной схемы, а именно $Ry = g$ (7.3), где искомое решение $y = (y_0^0,\, y_1^0,\, \ldots, y_N^0,\, y_0^1,\, y_1^1,\, \ldots, y_N^1,\, \ldots\ \ldots\ y_0^{K-1},\, y_1^{K-1},\, \ldots, y_N^{K-1},\, y_0^K,\, y_1^K,\, \ldots, y_N^K)^T$ и заданный вектор правых частей $g = (v_0,\, v_1,\, \ldots, v_N,\, \mu_1^1,\, \varphi_1^0,\, \varphi_2^0,\, \ldots, \varphi_{N-1}^0,\, \mu_2^1,$

$\mu_1^2,\, \varphi_1^1,\, \varphi_2^1,\, \ldots, \varphi_{N-1}^1,\, \mu_2^2,\, \ldots\ \ldots\ \mu_1^K,\, \varphi_1^{K-1},\, \varphi_2^{K-1},\, \ldots, \varphi_{N-1}^{K-1},\, \mu_2^K)^T$.

Представим вектор g в виде суммы $\bar{g} + \bar{\bar{g}}$. Вектор $\bar{g}$ получен из вектора g заменой $\varphi_j^n = 0\,(j = \overline{1,\, N-1},\, n = \overline{0, K-1})$. То есть $\bar{g} = (v_0,\, v_1,\, \ldots, v_N,\, \mu_1^1,\, 0,\, \ldots, 0,\, \mu_2^1,\, \mu_1^2,\, 0,\, \ldots, 0,\, \mu_2^2,\, \ldots\ \ldots\ \mu_1^K,\, 0,\, \ldots, 0,\, \mu_2^K)^T$. При замене $v_j = 0$, $\mu_1^n = 0$, $\mu_2^n = 0\,(j = \overline{0,\, N},\, n = \overline{1, K})$ в векторе g появляется вектор $\bar{\bar{g}} = (0,\, \ldots, 0,\, \varphi_1^0,\, \varphi_2^0,\, \ldots, \varphi_{N-1}^0,\, 0,\, 0,\, \varphi_1^1,\, \varphi_2^1,\, \ldots, \varphi_{N-1}^1,\, 0,\, \ldots$

$\ldots\ 0,\, \varphi_1^{K-1},\, \varphi_2^{K-1},\, \ldots, \varphi_{N-1}^{K-1},\, 0)^T$.

Решим две системы линейных алгебраических уравнений $R\bar{y} = \bar{g}$ и $R\bar{\bar{y}} = \bar{\bar{g}}$. Тогда сумма $\bar{y} + \bar{\bar{y}}$ будет решением разностной схемы $Ry = g$.

Вектор $\bar{y}$ равен $\bar{y} = (v_0,\, v_1,\, \ldots, v_N,\, \mu_1^1,\, 0,\, \ldots, 0,\, \mu_2^1,\, \mu_1^2,\, 0,\, \ldots, 0,\, \mu_2^2,\, \ldots\ \ldots\ \mu_1^K,\, 0,\, \ldots, 0,\, \mu_2^K)^T$. Компоненты вектора $\bar{y}$ удовлетворяют следующим неравенствам

$$|\bar{y}_j^n| \leqslant \max\left(\max_j |v_j|,\, \max_n |\mu_1^n|,\, \max_n |\mu_2^n|\right) \leqslant$$
$$\leqslant \max_j |v_j| + \max_n |\mu_1^n| + \max_n |\mu_2^n| = ||v||_h + ||\mu_1||_\tau + ||\mu_2||_\tau\,,$$

которые верны для любых $j = \overline{0,\, N}$ и $n = \overline{1, K}$. Следовательно,

$$\max_{n\,j} |\bar{y}_j^n| = ||\bar{y}||_{\tau\,h} \leqslant ||v||_h + ||\mu_1||_\tau + ||\mu_2||_\tau. \tag{7.6}$$

Компоненты вектора $\bar{\bar{y}} = (0,\, \ldots, 0,\, \bar{\bar{y}}_1^1,\, \bar{\bar{y}}_2^1,\, \ldots, \bar{\bar{y}}_{N-1}^1,\, 0,$

$0,\ \bar{\bar{y}}_1^2,\ \bar{\bar{y}}_2^2,\ \ldots,\bar{\bar{y}}_{N-1}^2,\ 0,\ \ldots\ \ldots\ 0,\ \bar{\bar{y}}_1^K,\ \bar{\bar{y}}_2^K,\ \ldots,\bar{\bar{y}}_{N-1}^K,\ 0)^T$ являются решением системы алгебраических уравнений

$$\frac{\bar{\bar{y}}_j^{n+1}-\bar{\bar{y}}_j^n}{\tau}=\frac{\bar{\bar{y}}_{j+1}^n-2\bar{\bar{y}}_j^n+\bar{\bar{y}}_{j-1}^n}{h^2}+\varphi_j^n,\ j=\overline{1,\,(N-1)},\ n=\overline{0,\,(K-1)}\,,\quad(7.7)$$

дополненной краевыми условиями $\bar{\bar{y}}_j^0=\bar{\bar{y}}_0^n=\bar{\bar{y}}_N^n=0,\ j=\overline{0,\,N},\ n=\overline{1,\,K}\,.$

Запишем уравнение (7.7) в виде

$$\bar{\bar{y}}_j^{n+1}=\left(1-\frac{2\tau}{h^2}\right)\bar{\bar{y}}_j^n+\frac{\tau}{h^2}(\bar{\bar{y}}_{j+1}^n+\bar{\bar{y}}_{j-1}^n)+\tau\varphi_j^n.$$

При любых допустимых значениях j и n верны следующие неравенства

$$\begin{aligned}|\bar{\bar{y}}_j^{n+1}|&\leqslant\left|1-\frac{2\tau}{h^2}\right|\left|\bar{\bar{y}}_j^n\right|+\frac{\tau}{h^2}\left(\left|\bar{\bar{y}}_{j+1}^n\right|+\left|\bar{\bar{y}}_{j-1}^n\right|\right)+\tau\left|\varphi_j^n\right|\leqslant\\&\leqslant\left(\left|1-\frac{2\tau}{h^2}\right|+\frac{2\tau}{h^2}\right)\max_j\left|\bar{\bar{y}}_j^n\right|+\tau\max_{n\,j}\left|\varphi_j^n\right|\ .\end{aligned}$$

Пусть выполнено неравенство

$$\frac{\tau}{h^2}\leqslant\frac{1}{2}.\qquad(7.8)$$

Тогда, раскрывая модуль, получим неравенство

$$|\bar{\bar{y}}_j^{n+1}|\leqslant\max_j\left|\bar{\bar{y}}_j^n\right|+\tau\left\|\varphi\right\|_{\tau\,h}\ ,$$

верное при любых допустимых значениях j и n. Здесь

$$\left\|\varphi\right\|_{\tau\,h}=\max_{n\,j}\left|\varphi_j^n\right|\ .$$

Применим это неравенство рекурсивно при получении оценки сверху для $|\bar{\bar{y}}_j^n|$:

$$\begin{aligned}|\bar{\bar{y}}_j^n|\leqslant\max\nolimits_j\left|\bar{\bar{y}}_j^{n-1}\right|+\tau\left\|\varphi\right\|_{\tau\,h}\leqslant\max\nolimits_j\left|\bar{\bar{y}}_j^{n-2}\right|+2\tau\left\|\varphi\right\|_{\tau\,h}\leqslant\ \ldots\ \leqslant\\\leqslant\max\nolimits_j\left|\bar{\bar{y}}_j^0\right|+n\tau\left\|\varphi\right\|_{\tau\,h}=\left\{\bar{\bar{y}}_j^0=0\right\}=n\tau\left\|\varphi\right\|_{\tau\,h}\leqslant\\\leqslant\left\{n\tau\leqslant T\leqslant c=constant\right\}\leqslant c\left\|\varphi\right\|_{\tau\,h}.\end{aligned}$$

Итак, при любых j и n верно неравенство $|\bar{\bar{y}}_j^n|\leqslant c\left\|\varphi\right\|_{\tau\,h}$. Следовательно

$$\max_{n\,j}\left|\bar{\bar{y}}_j^n\right|=\left\|\bar{\bar{y}}\right\|_{\tau\,h}\leqslant c\left\|\varphi\right\|_{\tau\,h}.$$

Тогда, для решения разностной схемы (7.2) верна оценка

$$\|y\|_{\tau h} = \|\bar{y} + \bar{\bar{y}}\|_{\tau h} \leqslant \|\bar{y}\|_{\tau h} + \|\bar{\bar{y}}\|_{\tau h} \leqslant ||v||_h + ||\mu_1||_\tau + ||\mu_2||_\tau + c\,\|\varphi\|_{\tau h}\,.$$

Выполнение этого неравенства означает, что разностная схема (7.2) устойчива по начальным, краевым условиям и по правой части. Так как неравенство верно при дополнительном условии (7.8), то явная разностная схема (7.2) называется условно устойчивой. Ограничение $\dfrac{\tau}{h^2} \leqslant \dfrac{1}{2}$ на параметры разностной схемы означает, что при использовании этой расчётной схемы шаг дискретной сетки по переменной t нужно брать достаточно мелким, порядка квадрата шага дискретной сетки по перменной x.

Итак, явная разностная схема (7.2) аппроксимирует дифференциальную задачу (7.1), имеет единственное решение и является условно устойчивой. Следовательно (см. Теорема (6.1)), решение разностной схемы сходится к точному решению дифференциальной задачи и имеет первый порядок точности по τ и второй порядок точности по h.

При исследовании свойств явной разностной схемы (7.2) наиболее трудоемким этапом было доказательство ее условной устойчивости. Математический аппарат, используемый для исследования произвольных линейных разностных схем на устойчивость, представлен в разделе «Теория устойчивости разностных схем» в учебной литературе по численным методам (см. [2], [3]). Эффективным способом получения необходимых условий устойчивости по начальным данным разностных схем для эволюционных задач является метод гармоник (см. [4]).

Метод гармоник.

Рассмотрим однородное разностное уравнение (7.2)

$$\frac{y_j^{n+1} - y_j^n}{\tau} = \frac{y_{j+1}^n - 2y_j^n + y_{j-1}^n}{h^2}, \tag{7.9}$$

с начальным условием частного вида $y_j^0 = e^{i\,jh\varphi}$, являющимся Фурье гармоникой. Здесь i — мнимая единица. Решение однородного уравнения с данным начальным условием имеет вид

$$y_j^n = q^n e^{i\,jh\varphi}, \tag{7.10}$$

где значение q определяется в результате подстановки (7.10) в уравнение (7.9).

Для выполнения условия устойчивости по начальным данным, то есть выполнения неравенства $\left|y_j^n\right| = \left|q^n e^{i\,jh\varphi}\right| \leqslant \left|y_j^0\right|$ при любых n, j и φ необходимо,

чтобы при всех вещественных φ выполнялось неравенство

$$|q| \leqslant 1, \tag{7.11}$$

являющееся необходимым условием устойчивости по начальным данным.

После подстановки (7.10) в (7.9) и деления всех слагаемых на общий сомножитель, получим

$$\frac{q-1}{\tau} = \frac{1}{h^2}\left(e^{ih\varphi} - 2 + e^{-ih\varphi}\right) \quad \Longrightarrow q = 1 - 4\frac{\tau}{h^2}\sin^2\frac{h\varphi}{2}.$$

То есть, неравенство (7.11) для явной разностной схемы (7.2) имеет вид

$$\left|1 - 4\frac{\tau}{h^2}\sin^2\frac{h\varphi}{2}\right| \leqslant 1,$$

или

$$-1 \leqslant 1 - 4\frac{\tau}{h^2}\sin^2\frac{h\varphi}{2} \leqslant 1.$$

Правое неравенство выполнено всегда, а левое неравенство приводится к виду

$$\frac{\tau}{h^2} \leqslant \frac{1}{2\sin^2\frac{h\varphi}{2}}.$$

Это неравенство должно выполняться при любых значениях φ. Следовательно, параметры τ и h явной разностной схемы должны удовлетворять неравенству

$$\frac{\tau}{h^2} \leqslant \frac{1}{2},$$

которое совпадает с ранее полученным условием устойчивости (7.8) явной разностной схемы.

7.1.2 Неявная разностная схема для уравнения теплопроводности

Используем для построения алгебраического аналога дифференциальной задачи (7.1) 4-ёх точечный шаблон, состоящий из трёх соседних узлов $(x_{j-1},\, t_{n+1})$, $(x_j,\, t_{n+1})$, $(x_{j+1},\, t_{n+1})$ дискретной сетки на $(n+1)$ —ом временном слое и одного дискретного узла $(x_j,\, t_n)$ на n —ом временном слое (см. Рис.7.2(а)). На этом шаблоне разностная схема для задачи (7.1) имеет вид

$$\begin{cases} \dfrac{y_j^{n+1}-y_j^n}{\tau} = \dfrac{y_{j+1}^{n+1}-2y_j^{n+1}+y_{j-1}^{n+1}}{h^2} + \varphi_j^{n+1}, \\ \qquad\qquad\qquad\qquad j=\overline{1,\,N-1},\ n=\overline{0,\,K-1}; \\ y_j^0 = v_j, \quad j=\overline{0,\,N}; \\ y_0^{n+1} = \mu_1^{n+1}, \quad n=\overline{0,\,K-1}; \\ y_N^{n+1} = \mu_2^{n+1}, \quad n=\overline{0,\,K-1}. \end{cases} \tag{7.12}$$

Во внутренних узлах дискретной сетки невязка в разностной схеме (7.12) равна

$$\psi_j^{n+1} = -\frac{u_j^{n+1}-u_j^n}{\tau} + \frac{1}{h^2}\left(u_{j+1}^{n+1}-2u_j^{n+1}+u_{j-1}^{n+1}\right) + \varphi_j^{n+1}.$$

Представим $u_{j\pm1}^{n+1}$ и u_j^n в виде разложений в ряд Тейлора

$$\begin{aligned} u_{j\pm1}^{n+1} &= u_j^{n+1} \pm u_{x,j}^{n+1}h + u_{xx,j}^{n+1}\frac{h^2}{2} \pm u_{xxx,j}^{n+1}\frac{h^3}{6} + O(h^4); \\ u_j^n &= u_j^{n+1} - u_{t,j}^{n+1}\tau + O(\tau^2) \end{aligned}$$

и подставим в выражение для невязки. Тогда невязка примет вид

$$\psi_j^{n+1} = (-u_{t,j}^{n+1} + u_{xx,j}^{n+1} + f_j^{n+1}) + \varphi_j^{n+1} - f_j^{n+1} + O(\tau + h^2).$$

Так как u_j^{n+1} является решением уравнения теплопроводности (7.1), то комбинация в скобках в правой части равна нулю. Если для правой части разностной схемы (7.12) выполнено условие $\varphi_j^{n+1} = f_j^{n+1} + O(\tau + h^2)$, то невязка равна $\psi_j^{n+1} = O(\tau + h^2)$, $j=\overline{1,\,N-1}$, $n=\overline{0, K-1}$. Следовательно,

$$||\psi||_{\tau,h} = \max_{j,n}|\psi_j^n| = O(\tau + h^2).$$

Итак, разностная схема (7.12) аппроксимирует краевую задачу (7.1) с первым порядком по τ и вторым порядком по h.

Рассмотрим вопрос о методе решения разностной схемы. Запишем разностную схему (7.12) в следующем виде

$$\begin{cases} \dfrac{\tau}{h^2}y_{j-1}^{n+1} - \left(1+\dfrac{2\tau}{h^2}\right)y_j^{n+1} + \dfrac{\tau}{h^2}y_{j+1}^{n+1} = -y_j^n - \tau\varphi_j^{n+1}, \\ \qquad\qquad\qquad\qquad j=\overline{1,\,N-1},\ n=\overline{0,\,K-1}; \\ y_j^0 = v_j, \quad j=\overline{0,\,N}; \\ y_0^{n+1} = \mu_1^{n+1}, \quad n=\overline{0,\,K-1}; \\ y_N^{n+1} = \mu_2^{n+1}, \quad n=\overline{0,\,K-1}. \end{cases}$$

Введём обозначения:

$$A = \frac{\tau}{h^2}, \quad B = \frac{\tau}{h^2}, \quad C = \left(1 + \frac{2\tau}{h^2}\right), \quad F_j^n = y_j^n + \tau\varphi_j^{n+1}.$$

Тогда, разностная схема примет вид

$$\begin{cases} Ay_{j-1}^{n+1} - Cy_j^{n+1} + By_{j+1}^{n+1} = -F_j^n,\ j = \overline{1,\ N-1},\ n = \overline{0,\ K-1}; \\ y_j^0 = v_j, \quad j = \overline{0,\ N}; \\ y_0^{n+1} = \mu_1^{n+1}, \quad n = \overline{0,\ K-1}; \\ y_N^{n+1} = \mu_2^{n+1}, \quad n = \overline{0,\ K-1}. \end{cases} \tag{7.13}$$

Запишем (7.13) в частном случае, когда $n = 0$:

$$\begin{cases} Ay_{j-1}^1 - Cy_j^1 + By_{j+1}^1 = -F_j^0, \quad j = \overline{1,\ N-1}; \\ y_0^1 = \mu_1^1, \quad y_N^1 = \mu_2^1; \\ F_j^0 = v_j + \tau\varphi_j^1, \quad j = \overline{1,\ N-1}. \end{cases} \tag{7.14}$$

Соотношения (7.14) являются системой линейных алгебраических уравнений относительно неизвестных $y_0^1,\ y_1^1,\ \ldots,\ y_N^1$. Эффективным методом решения системы (7.14) является метод прогонки.

Решив систему (7.14), запишем (7.13) в случае, когда $n = 1$:

$$\begin{cases} Ay_{j-1}^2 - Cy_j^2 + By_{j+1}^2 = -F_j^1, \quad j = \overline{1,\ N-1}; \\ y_0^2 = \mu_1^2, \quad y_N^2 = \mu_2^2; \\ F_j^1 = y_j^1 + \tau\varphi_j^2, \quad j = \overline{1,\ N-1}. \end{cases} \tag{7.15}$$

Система (7.15) является системой линейных алгебраических уравнений относительно неизвестных $y_0^2,\ y_1^2,\ \ldots,\ y_N^2$. Решается эта системы методом прогонки.

Последовательно перебирая все значения $n = \overline{0,\ K-1}$ в (7.13), вычисляются y_j^n на всех временных слоях. Так как для определения y_j^n на каждом временном слое необходимо решать методом прогонки систему линейных алгебраических уравнений, то разностную схему (7.13) называют неявной разностной схемой.

Для исследования устойчивости разностной схемы (7.12) используем метод гармоник. Сопоставим разностной схеме однородное уравнение

$$\frac{y_j^{n+1} - y_j^n}{\tau} = \frac{y_{j+1}^{n+1} - 2y_j^{n+1} + y_{j-1}^{n+1}}{h^2}.$$

Подставим в однородное уравнение решение частного вида $y_j^n = q^n e^{i\,jh\varphi}$. После деления всех слагаемых на общий сомножитель, получим уравнение относительно параметра q:

$$\frac{q-1}{\tau} = q\frac{e^{ih\varphi} - 2 + e^{-ih\varphi}}{h^2} \quad \Longrightarrow \quad \frac{q-1}{\tau} = -q\frac{4}{h^2}\sin^2\frac{h\varphi}{2}.$$

Отсюда следует, что $q = \dfrac{1}{\left(1 + \dfrac{4\tau}{h^2}\sin^2\frac{h\varphi}{2}\right)}$. Так как знаменатель дроби больше или равен единице, то условие устойчивости $|q| \leqslant 1$ выполнено для любых допустимых значений параметров τ и h.

Итак, ограничений на τ и h нет, поэтому неявная разностная схема (7.12) является абсолютно устойчивой. Абсолютная устойчивость является преимуществом неявных разностных схем, так как величина шагов дискретной сетки τ и h в абсолютно устойчивых разностных схемах определяется только необходимой точностью расчёта, а не особенностью разностной схемы.

7.1.3 Разностная схема с весами для уравнения теплопроводности

Используем для построения разностной схемы для дифференциальной задачи (7.1) 6-ти точечный шаблон, состоящий из трёх соседних узлов

$(x_{j-1},\, t_{n+1})$, $(x_j,\, t_{n+1})$, $(x_{j+1},\, t_{n+1})$ дискретной сетки на $(n+1)$ —ом временном слое и трёх соседних узлов $(x_{j-1},\, t_n)$, $(x_j,\, t_n)$, $(x_{j+1},\, t_n)$ на n —ом временном слое (см. Рис.7.2(б)). Зададим произвольный действительный параметр $\sigma \in [0;\, 1]$ — весовой множитель для второй разностной производной. На этом шаблоне для задачи (7.1) построим разностную схему вида

$$\begin{cases} \dfrac{y_j^{n+1} - y_j^n}{\tau} = \sigma\dfrac{y_{j+1}^{n+1} - 2y_j^{n+1} + y_{j-1}^{n+1}}{h^2} + (1-\sigma)\dfrac{y_{j+1}^n - 2y_j^n + y_{j-1}^n}{h^2} + \varphi_j^n, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad j = \overline{1,\, N-1}, \quad n = \overline{0,\, K-1}; \\ y_j^0 = v_j, \quad j = \overline{0,\, N}; \\ y_0^{n+1} = \mu_1^{n+1}, \quad n = \overline{0,\, K-1}; \\ y_N^{n+1} = \mu_2^{n+1}, \quad n = \overline{0,\, K-1}. \end{cases} \tag{7.16}$$

Систему уравнений (7.16) называют разностной схемой с весами для уравнения теплопроводности.

Точка $(x_j,\, t_{n+\frac{1}{2}})$ является центром симметрии для 6-ти точечного шаблона. Отнесённая к этой точке невязка в разностной схеме (7.16) равна

$$\psi_j^{n+\frac{1}{2}} = -\frac{u_j^{n+1} - u_j^n}{\tau} + \sigma u_{\overline{x}x,j}^{n+1} + (1-\sigma) u_{\overline{x}x,j}^n + \varphi_j^n.$$

Представим u_j^{n+1} и u_j^n в виде разложений в ряд Тейлора

$$\begin{aligned} u_j^{n+1} = u\left(x_j, t_{n+\frac{1}{2}} + \frac{\tau}{2}\right) &= u_j^{n+\frac{1}{2}} + u_{t,j}^{n+\frac{1}{2}} \frac{\tau}{2} + u_{tt,j}^{n+\frac{1}{2}} \frac{1}{2} \left(\frac{\tau}{2}\right)^2 + O(\tau^3), \\ u_j^n = u\left(x_j, t_{n+\frac{1}{2}} - \frac{\tau}{2}\right) &= u_j^{n+\frac{1}{2}} - u_{t,j}^{n+\frac{1}{2}} \frac{\tau}{2} + u_{tt,j}^{n+\frac{1}{2}} \frac{1}{2} \left(\frac{\tau}{2}\right)^2 + O(\tau^3) \end{aligned}$$

и подставим эти разложения в невязку. Тогда получим

$$\psi_j^{n+\frac{1}{2}} = -u_{t,j}^{n+\frac{1}{2}} + O(\tau^2) + \sigma u_{\overline{x}x,j}^{n+1} + (1-\sigma) u_{\overline{x}x,j}^n + \varphi_j^n. \tag{7.17}$$

В точке $(x_j,\, t)$, где t любое, вторая разностная производная равна

$$u_{\overline{x}x}(x_j, t) = \frac{1}{h^2} \left(u(x_{j+1}, t) - 2u(x_j, t) + u(x_{j-1}, t)\right).$$

Представим $u(x_{j\pm1},\, t)$, входящие в выражение для второй разностной производной, в виде разложения в ряд Тейлора

$$\begin{aligned} u(x_{j\pm1}, t) = u(x_j, t) \pm u_x(x_j, t)h + u_{xx}(x_j, t)\frac{h^2}{2} \pm u_{xxx}(x_j, t)\frac{h^3}{3!} + \\ + u_{xxxx}(x_j, t)\frac{h^4}{4!} \pm u_{xxxxx}(x_j, t)\frac{h^5}{5!} + O(h^6). \end{aligned}$$

Тогда, вторая разностная производная имеет вид

$$u_{\overline{x}x}(x_j, t) = u_{xx}(x_j, t) + u_{xxxx}(x_j, t)\frac{h^2}{12} + O(h^4).$$

Используем полученное соотношение при разложении вторых разностных производных $u_{\overline{x}x,j}^{n+1}$ и $u_{\overline{x}x,j}^n$, входящих в (7.17), в ряд Тейлора с центром в точке $(x_j,\, t_{n+\frac{1}{2}})$.

Верны следующие преобразования:

$$u^{n+1}_{\overline{x}x,j} = u_{\overline{x}x}(x_j,\, t_{n+1}) = u_{\overline{x}x}\left(x_j,\, t_{n+\frac{1}{2}} + \frac{\tau}{2}\right) = u_{xx}\left(x_j,\, t_{n+\frac{1}{2}} + \frac{\tau}{2}\right) + \\ + u_{xxxx}\left(x_j,\, t_{n+\frac{1}{2}} + \frac{\tau}{2}\right)\frac{h^2}{12} + O(h^4) = u^{n+\frac{1}{2}}_{xx,j} + u^{n+\frac{1}{2}}_{xxt,j}\frac{\tau}{2} + \\ + u^{n+\frac{1}{2}}_{xxxx,j}\frac{h^2}{12} + u^{n+\frac{1}{2}}_{xxxxt,j}\frac{\tau}{2}\frac{h^2}{12} + O(\tau^2 + h^4);$$

$$u^{n}_{\overline{x}x,j} = u_{\overline{x}x}(x_j,\, t_{n}) = u_{\overline{x}x}\left(x_j,\, t_{n+\frac{1}{2}} - \frac{\tau}{2}\right) = u_{xx}\left(x_j,\, t_{n+\frac{1}{2}} - \frac{\tau}{2}\right) + \\ + u_{xxxx}\left(x_j,\, t_{n+\frac{1}{2}} - \frac{\tau}{2}\right)\frac{h^2}{12} + O(h^4) = u^{n+\frac{1}{2}}_{xx,j} - u^{n+\frac{1}{2}}_{xxt,j}\frac{\tau}{2} + \\ + u^{n+\frac{1}{2}}_{xxxx,j}\frac{h^2}{12} - u^{n+\frac{1}{2}}_{xxxxt,j}\frac{\tau}{2}\frac{h^2}{12} + O(\tau^2 + h^4).$$

Подставим полученные соотношения в (7.17). Тогда, выражение для невязки примет вид

$$\psi^{n+\frac{1}{2}}_j = \left(-u^{n+\frac{1}{2}}_{t,j} + u^{n+\frac{1}{2}}_{xx,j} + f^{n+\frac{1}{2}}_j\right) + \varphi^n_j - f^{n+\frac{1}{2}}_j + u^{n+\frac{1}{2}}_{xxxx,j}\frac{h^2}{12} + \left(\sigma - \frac{1}{2}\right)u^{n+\frac{1}{2}}_{xxt,j}\tau + \\ + \left(\sigma - \frac{1}{2}\right)u^{n+\frac{1}{2}}_{xxxxt,j}\tau\frac{h^2}{12} + O(\tau^2 + h^4).$$

Функция $u(x,\, t)$ является решением уравнения теплопроводности. Поэтому, в правой части первая комбинация в скобках обращается в ноль и невязка равна

$$\psi^{n+\frac{1}{2}}_j = \varphi^n_j - f^{n+\frac{1}{2}}_j + u^{n+\frac{1}{2}}_{xxxx,j}\frac{h^2}{12} + \left(\sigma - \frac{1}{2}\right)u^{n+\frac{1}{2}}_{xxt,j}\tau + \\ + \left(\sigma - \frac{1}{2}\right)u^{n+\frac{1}{2}}_{xxxxt,j}\tau\frac{h^2}{12} + O(\tau^2 + h^4). \tag{7.18}$$

Рассмотрим два частных случая. Пусть в разностной схеме (7.16) параметр $\sigma = \frac{1}{2}$. В этом случае разностную схему называют симметричной разностной схемой и невязка (7.18) имеет вид

$$\psi^{n+\frac{1}{2}}_j = \varphi^n_j - f^{n+\frac{1}{2}}_j + u^{n+\frac{1}{2}}_{xxxx,j}\frac{h^2}{12} + O(\tau^2 + h^4).$$

Если при вычислении правой части в разностной схеме (7.16) выполнено условие $\varphi^n_j = f^{n+\frac{1}{2}}_j + O(\tau^2 + h^2)$, то невязка

$$\psi^{n+\frac{1}{2}}_j = O(\tau^2 + h^2).$$

Следовательно, симметричная разностная схема аппроксимирует дифференциальную задачу (7.1) с 2-ым порядком аппроксимации по τ и h.

Второй частный случай. Функция $u(x,\,t)$ является решением уравнения теплопроводности $u_t = u_{xx} + f$. Поэтому, в случае достаточной гладкости функций $u(x,\,t)$ и $f(x,\,t)$, для четвёртой производной верно равенство $u_{xxxx} = u_{xxt} - f_{xx}$.

Тогда формула (7.18) для невязки принимает вид

$$\psi_j^{n+\frac{1}{2}} = \varphi_j^n - f_j^{n+\frac{1}{2}} - f_{xx,j}^{n+\frac{1}{2}}\,\frac{h^2}{12} + \left(\left(\sigma - \frac{1}{2}\right)\tau + \frac{h^2}{12}\right) u_{xxt,j}^{n+\frac{1}{2}} + \\ + \left(\sigma - \frac{1}{2}\right) u_{xxxxt,j}^{n+\frac{1}{2}}\,\tau\frac{h^2}{12} + O(\tau^2 + h^4).$$

Пусть $\sigma = \sigma^* = \dfrac{1}{2} - \dfrac{h^2}{12\tau}$ и $\varphi_j^n = f_j^{n+\frac{1}{2}} + f_{xx,j}^{n+\frac{1}{2}}\,\dfrac{h^2}{12} + O(\tau^2 + h^4)$.

Тогда невязка будет равна

$$\psi_j^{n+\frac{1}{2}} = O(\tau^2 + h^4).$$

То есть, разностная схема с $\sigma = \sigma^*$ аппроксимирует дифференциальную задачу (7.1) с 2-ым порядком по τ и 4-ым порядком аппроксимации по h. Такая разностная схема называется разностной схемой повышенного порядка аппроксимации.

При всех других значениях $\sigma \in [0;\,1]$ $\left(\sigma \neq \dfrac{1}{2},\, \sigma \neq \sigma^*\right)$ и $\varphi_j^n = f_j^{n+\frac{1}{2}} + O(\tau + h^2)$ невязка равна

$$\psi_j^{n+\frac{1}{2}} = O(\tau + h^2)$$

и разностная схема (7.16) аппроксимирует дифференциальную задачу (7.1) с 1-ым порядком по τ и 2-ым порядком аппроксимации по h.

Рассмотрим вопрос о методе решения разностной схемы. Запишем разностную схему (7.16) в виде

$$\begin{cases} \sigma\dfrac{\tau}{h^2}y_{j-1}^{n+1} - \left(1 + \sigma\dfrac{2\tau}{h^2}\right) y_j^{n+1} + \sigma\dfrac{\tau}{h^2}y_{j+1}^{n+1} = -y_j^n - (1-\sigma)\tau y_{\overline{x}x,j}^n - \tau\varphi_j^n, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad j = \overline{1,\,N-1},\, n = \overline{0,\,K-1}; \\ y_j^0 = v_j, \quad j = \overline{0,\,N}; \\ y_0^{n+1} = \mu_1^{n+1}, \quad n = \overline{0,\,K-1}; \\ y_N^{n+1} = \mu_2^{n+1}, \quad n = \overline{0,\,K-1}. \end{cases}$$

Введём обозначения:

$$A = \sigma\frac{\tau}{h^2}, \quad B = \sigma\frac{\tau}{h^2}, \quad C = \left(1 + \sigma\frac{2\tau}{h^2}\right), \quad F_j^n = y_j^n + (1-\sigma)\tau y_{\overline{x}x,j}^n + \tau\varphi_j^n.$$

Тогда, разностная схема примет вид, совпадающий с (7.13)

$$\begin{cases} Ay_{j-1}^{n+1} - Cy_j^{n+1} + By_{j+1}^{n+1} = -F_j^n,\ j = \overline{1,\ (N-1)},\ n = \overline{0,\ (K-1)}; \\ y_j^0 = v_j, \quad j = \overline{0,\ N}; \\ y_0^{n+1} = \mu_1^{n+1}, \quad n = \overline{0,\ (K-1)}; \\ y_N^{n+1} = \mu_2^{n+1}, \quad n = \overline{0,\ (K-1)}, \end{cases} \tag{7.19}$$

где $F_j^n = y_j^n + (1-\sigma)\tau y_{\overline{x}x,j}^n + \tau\varphi_j^n,\ j = \overline{1,\ (N-1)},\ n = \overline{0,\ (K-1)}$.

Повторим то, что делали при решении уравнений (7.13). Запишем систему (7.19) в частном случае, когда $n = 0$:

$$\begin{cases} Ay_{j-1}^1 - Cy_j^1 + By_{j+1}^1 = -F_j^0, \quad j = \overline{1,\ (N-1)}; \\ y_0^1 = \mu_1^1, \quad y_N^1 = \mu_2^1; \\ F_j^0 = v_j + (1-\sigma)\tau v_{\overline{x}x,j} + \tau\varphi_j^0, \quad j = \overline{1,\ (N-1)}. \end{cases}$$

Получили систему линейных алгебраических уравнений относительно неизвестных $y_0^1,\ y_1^1,\ \ldots,\ y_N^1$. Эффективным методом решения данной системы является метод прогонки. Решив эту систему, запишем (7.19) при $n = 1$. Получим систему линейных алгебраических уравнений относительно неизвестных $y_0^2,\ y_1^2,\ \ldots,\ y_N^2$. Решается эта системы методом прогонки. Последовательно перебирая все значения $n = \overline{0,\ K-1}$ в (7.19), вычисляются y_j^n на всех временных слоях.

Для исследования устойчивости разностной схемы (7.16) используем метод гармоник. Сопоставим разностной схеме однородное уравнение

$$\frac{y_j^{n+1} - y_j^n}{\tau} = \sigma\frac{y_{j+1}^{n+1} - 2y_j^{n+1} + y_{j-1}^{n+1}}{h^2} + (1-\sigma)\frac{y_{j+1}^n - 2y_j^n + y_{j-1}^n}{h^2}.$$

Подставим в однородное уравнение решение частного вида $y_j^n = q^n e^{i\,jh\varphi}$. После деления всех слагаемых на общий сомножитель, получим уравнение относительно параметра q:

$$\frac{q-1}{\tau} = \sigma\frac{q}{h^2}(e^{ih\varphi} - 2 + e^{-ih\varphi}) + (1-\sigma)\frac{1}{h^2}(e^{ih\varphi} - 2 + e^{-ih\varphi}) \iff$$
$$\iff \frac{q-1}{\tau} = -4\sigma\frac{q}{h^2}\sin^2\frac{h\varphi}{2} - 4(1-\sigma)\frac{1}{h^2}\sin^2\frac{h\varphi}{2}.$$

Решением этого уравнения является

$$q = \frac{1 - (1-\sigma)4\dfrac{\tau}{h^2}\sin^2\dfrac{h\varphi}{2}}{1 + 4\sigma\dfrac{\tau}{h^2}\sin^2\dfrac{h\varphi}{2}}.$$

Условие устойчивости разностной схемы $|q| \leqslant 1$ в данном случае приводит к системе двух неравенств

$$\begin{cases} 1-(1-\sigma)4\dfrac{\tau}{h^2}\sin^2\dfrac{h\varphi}{2} \;\leqslant\; 1+4\sigma\dfrac{\tau}{h^2}\sin^2\dfrac{h\varphi}{2}; \\ -1-4\sigma\dfrac{\tau}{h^2}\sin^2\dfrac{h\varphi}{2} \;\leqslant\; 1-(1-\sigma)4\dfrac{\tau}{h^2}\sin^2\dfrac{h\varphi}{2}. \end{cases}$$

Первое неравенство в системе выполнено всегда. Второе неравенство приводится к виду

$$-\sigma 8\frac{\tau}{h^2}\sin^2\frac{h\varphi}{2} \;\leqslant\; -4\frac{\tau}{h^2}\sin^2\frac{h\varphi}{2}+2 \quad\Longleftrightarrow\quad \sigma \geqslant \frac{1}{2}-\frac{h^2}{4\tau\sin^2\dfrac{h\varphi}{2}}.$$

Следовательно, условие устойчивости $|q| \leqslant 1$ будет выполняться при любом φ, если весовой множитель σ удовлетворяет неравенству

$$\sigma \geqslant \frac{1}{2}-\frac{h^2}{4\tau}.$$

Заметим, что симметричная разностная схема с весовым множителем $\sigma = \dfrac{1}{2}$ является абсолютно устойчивой разностной схемой.

Итак, на примере дифференциальной краевой задачи (7.1), имеющей аналитическое решение, были показаны методы построения разностных схем, применимые и для краевых задач, для которых аналитические решения не известны.

7.1.4 Разностная схема для уравнения теплопроводности с переменными коэффициентами

Рассмотрим краевую задачу для уравнения теплопроводности с переменными коэффициентами

$$\begin{cases} \rho(x,t)\dfrac{\partial u}{\partial t} = \dfrac{\partial}{\partial x}\left(k(x,t)\dfrac{\partial u}{\partial x}\right)+f(x,t), \quad 0<x<l,\ 0<t; \\ u(0,t)=\mu_1(t), \quad u(l,t)=\mu_2(t), \quad 0<t; \\ u(x,0)=v(x), \quad 0\leqslant x\leqslant l. \end{cases} \tag{7.20}$$

Здесь $\rho(x,t),\ k(x,t),\ f(x,t)$ гладкие функции, удовлетворяющие условиям

$$0<c_1\leqslant\rho(x,t),\ 0<c_2\leqslant k(x,t)\leqslant c_3,$$

при которых решение краевой задачи существует и единственно.

Производной $\dfrac{\partial u(x,t)}{\partial t}$ на отрезке $t \in [t_n;\, t_{n+1}]$ сопоставим первую разностную производную

$$\frac{\partial u(x,t)}{\partial t} \approx \frac{u(x,t_{n+1}) - u(x,t_n)}{\tau}.$$

Производной $\dfrac{\partial}{\partial x}\left(k(x,t)\dfrac{\partial u}{\partial x}\right)\Big|_{x=x_j}$ сопоставим алгебраическую комбинацию, полученную с использованием интегро-интерполяционного метода, следующего вида

$$\frac{\partial}{\partial x}\left(k(x,t)\frac{\partial u}{\partial x}\right)\Big|_{x=x_j} \approx \left(au_{\overline{x}}\right)_{x,j} = \frac{1}{h}\left(a_{j+1}\frac{u(x_{j+1},t)-u(x_j,t)}{h} - a_j\frac{u(x_j,t)-u(x_{j-1},t)}{h}\right),$$

где a_j вычисляются по формуле

$$a_j = \left(\frac{1}{h}\int\limits_{x_{j-1}}^{x_j}\frac{dx}{k(x,t)}\right)^{-1} \approx k(x_{j-\frac{1}{2}},t).$$

Используем шаблон из шести дискретных точек (Рис.7.2(б)). На этом шаблоне для краевой задачи (7.20) построим разностную схему с весами

$$\begin{cases} \rho(x_j,t)\,\dfrac{y_j^{n+1}-y_j^n}{\tau} = \sigma\left(ay_{\overline{x}}\right)_{x,j}^{n+1} + (1-\sigma)\left(ay_{\overline{x}}\right)_{x,j}^{n} + \varphi_j^n, \\ \qquad\qquad j = \overline{1,\, N-1},\ n = \overline{0,\, K-1}; \\ y_0^{n+1} = \mu_1^{n+1}, \quad y_N^{n+1} = \mu_2^{n+1}, \quad n = \overline{0,\, K-1}; \\ y_j^0 = v_j, \quad j = \overline{0,\, N}. \end{cases} \tag{7.21}$$

В разностной схеме (7.21) при вычислении значений коэффициентов $\rho(x_j,t)$ и a_j можно выбрать любое значение $t \in [t_n;\, t_{n+1}]$. Схема (7.21) имеет второй порядок аппроксимации по τ и по h, если взять $t = t_{n+\frac{1}{2}}$, $\varphi_j^n = f_j^{n+\frac{1}{2}} + O(\tau^2+h^2)$ и $\sigma = \dfrac{1}{2}$. При других значениях σ, t и $\varphi_j^n = f_j^{n+\frac{1}{2}} + O(\tau+h^2)$ схема имеет первый порядок аппроксимации по τ и второй порядок по h.

Схема с переменными коэффициентами (7.21) абсолютно устойчива при $\sigma \geqslant \dfrac{1}{2}$.

Решение разностной схемы (7.21) проводится по временным слоям. Последовательно перебирая все значения $n = \overline{0,\, K-1}$ в (7.21), вычисляются y_j^n на всех временных слоях. На каждом временном слое необходимо решать методом прогонки систему линейных алгебраических уравнений.

7.1.5 Разностная схема для нелинейного уравнения теплопроводности

Рассмотрим краевую задачу для нелинейного уравнения теплопроводности

$$\begin{cases} \dfrac{\partial u}{\partial t} = \dfrac{\partial}{\partial x}\left(k(u,x,t)\dfrac{\partial u}{\partial x}\right) + f(u,x,t), \quad 0 < x < l,\ 0 < t; \\ u(0,t) = \mu_1(t), \quad u(l,t) = \mu_2(t), \quad 0 < t; \\ u(x,0) = v(x), \quad 0 \leqslant x \leqslant l. \end{cases} \tag{7.22}$$

Коэффициент теплопроводности $k(u,x,t)$ в этом уравнении зависит от искомого решения $u(x,t)$ и пределы изменения его значений заранее не известны. При численном решении нелинейных дифференциальных уравнений используются неявные абсолютно устойчивые разностные схемы.

Сопоставим краевой задаче (7.22) неявную разностную схему

$$\begin{cases} \dfrac{y_j^{n+1} - y_j^n}{\tau} = \left(a(y_j^{n+1})y_{\overline{x}}^{n+1}\right)_{x,j} + \varphi(y_j^{n+1}),\ j = \overline{1,\ (N-1)},\ n = \overline{0,\ (K-1)}; \\ y_0^{n+1} = \mu_1^{n+1}, \quad y_N^{n+1} = \mu_2^{n+1}, \quad n = \overline{0,\ (K-1)}; \\ y_j^0 = v_j, \quad j = \overline{0,\ N}, \end{cases} \tag{7.23}$$

где $a(y_j^{n+1}) = k(y_j^{n+1}, x_{j-\frac{1}{2}}, t_{n+1})$ и $\varphi(y_j^{n+1}) = f(y_j^{n+1}, x_j, t_{n+1})$.

Разностная схема (7.23) представляет собой систему нелинейных алгебраических уравнений. Раскрыв в разностном уравнении выражение для второй разностной производной, получим следующее нелинейное уравнение

$$\frac{y_j^{n+1} - y_j^n}{\tau} = \frac{1}{h}\left(a(y_{j+1}^{n+1})\,\frac{y_{j+1}^{n+1} - y_j^{n+1}}{h} - a(y_j^{n+1})\,\frac{y_j^{n+1} - y_{j-1}^{n+1}}{h}\right) + \varphi(y_j^{n+1}).$$

Для получения решения y_j^{n+1} этого нелинейного уравнения необходимо на каждом временном слое использовать итерационный метод, например, следующего вида

$$\frac{\overset{(k+1)}{y_j} - y_j^n}{\tau} = \frac{1}{h}\left(a(\overset{(k)}{y_{j+1}})\,\frac{\overset{(k+1)}{y_{j+1}} - \overset{(k+1)}{y_j}}{h} - a(\overset{(k)}{y_j})\,\frac{\overset{(k+1)}{y_j} - \overset{(k+1)}{y_{j-1}}}{h}\right) + \varphi(\overset{(k)}{y_j}),$$

где $\overset{(k+1)}{y_j}$ это $(k+1)$-ое итерационное приближение для y_j^{n+1}, k- номер итерации, $k = \overline{0,\ M-1}$, $(M \leqslant 5)$, $y_j^{n+1} = \overset{(M)}{y_j}$ и $\overset{(0)}{y_j} = y_j^n$. Значения итерационного приближения $\overset{(k+1)}{y_j}$ при $j = \overline{1,\ N-1}$ находятся методом прогонки, а $\overset{(k+1)}{y_0} = \mu_1^{n+1}$ и $\overset{(k+1)}{y_N} = \mu_2^{n+1}$.

7.2 Разностная схема для уравнения колебаний

Рассмотрим краевую задачу для одномерного по пространству уравнения колебаний

$$\begin{cases} u_{tt} = u_{xx} + f(x,t), \quad 0 < x < l,\ 0 < t; \\ u(0,t) = \mu_1(t), \quad u(l,t) = \mu_2(t), \quad 0 < t; \\ u(x,0) = v(x), \quad u_t(x,0) = w(x), \quad 0 \leqslant x \leqslant l. \end{cases} \tag{7.24}$$

Решение данной задачи существует и единственно.

Введем на ограниченной части допустимых значений переменных $0 \leqslant t \leqslant T$ и $0 \leqslant x \leqslant l$ дискретную сетку, состоящую из узлов, являющихся точками пересечения двух семейств линий $x = x_j = jh,\ j = \overline{0,\ N},\ h = \frac{l}{N}$ и $t = t_n = n\tau,\ n = \overline{0,\ K},\ \tau = \frac{T}{K}$.

В условиях задачи (7.24) присутствуют производные u_{tt}, u_{xx} и u_t.

Значения производных u_{tt} и u_{xx} приближённо заменим на вторые разностные производные

$$u_{tt}(x,t_n) = u_{\bar{t}t}(x,t_n) + O(\tau^2) = \frac{1}{\tau^2}\left(u(x,t_{n+1}) - 2u(x,t_n) + u(x,t_{n-1})\right) + O(\tau^2);$$

$$u_{xx}(x_j,t) = u_{\bar{x}x}(x_j,t) + O(h^2) = \frac{1}{h^2}\left(u(x_{j+1},t) - 2u(x_j,t) + u(x_{j-1},t)\right) + O(h^2).$$

Первую разностную производную по t в начальный момент времени $t = 0$ представим в виде

$$\frac{u(x,\tau) - u(x,0)}{\tau} = \frac{1}{\tau}\left(u(x,0) + u_t(x,0)\tau + u_{tt}(x,0)\frac{\tau^2}{2} + O(\tau^3) - u(x,0)\right) =$$

$$= u_t(x,0) + u_{tt}(x,0)\frac{\tau}{2} + O(\tau^2).$$

Пусть уравнение колебаний верно в момент времени $t = 0$. Тогда $u_{tt}(x,0) = u_{xx}(x,0) + f(x,0)$ и выражение для первой разностной производной по t, с учётом начальных условий задачи (7.24), примет вид

$$\frac{u(x,t_1) - u(x,t_0)}{\tau} = w(x) + \left(v_{xx}(x) + f(x,0)\right)\frac{\tau}{2} + O(\tau^2).$$

Сопоставим краевой задаче (7.24) следующую разностную схему

$$\begin{cases} y^n_{\bar{t}t,j} = y^n_{\bar{x}x,j} + \varphi^n_j, \quad j = \overline{1,\ N-1},\ n = \overline{1,\ K-1}; \\ y^n_0 = \mu^n_1, \quad y^n_N = \mu^n_2, \quad n = \overline{1,\ K}; \\ y^0_j = v_j, \quad \dfrac{y^1_j - y^0_j}{\tau} = w_j + \left(v_{xx,j} + f^0_j\right)\dfrac{\tau}{2}, \quad j = \overline{0,\ N}, \end{cases} \tag{7.25}$$

где $f_j^0 = f(x_j, 0)$ и $\varphi_j^n = f(x_j, t_n) + O(h^2 + \tau^2)$.

Разностная схема (7.25) построена на 5-ти точечном шаблоне (см. Рис.7.4) и аппроксимирует краевую задачу (7.24) с 2-ым порядком по τ и h, то есть $\psi_j^n = O(h^2 + \tau^2)$.

Рассмотрим вопрос о методе решения разностной схемы. Два начальных условия в разностной схеме (7.25) определяют искомую сеточную функцию y_j^n на временных слоях с $n = 0$ и $n = 1$:

$$y_j^0 = v_j;$$
$$y_j^1 = v_j + \tau\left(w_j + \left(v_{xx,j} + f_j^0\right)\frac{\tau}{2}\right), \quad j = \overline{0,\, N}.$$

Раскрыв выражение для второй разностной производной по t, запишем разностное уравнение схемы (7.25) в виде

$$y_j^{n+1} = 2y_j^n - y_j^{n-1} + \tau^2 y_{\overline{x}x,j}^n + \tau^2\varphi_j^n \tag{7.26}$$

и будем считать это соотношение уравнением относительно y_j^{n+1}.

При $n = 1$ уравнение (7.26) принимает вид

$$y_j^2 = 2y_j^1 - y_j^0 + \frac{\tau^2}{h^2}\left(y_{j+1}^1 - 2y_j^1 + y_{j-1}^1\right) + \tau^2\varphi_j^n, \quad j = \overline{0,\, N}.$$

Так как y_j^0 и y_j^1 уже известны, то можем найти y_j^2. Аналогично вычисляются y_j^n на третьем и всех последующих временных слоях.

Таким образом, данная разностная схема решается по временным слоям и решение разностной схемы (7.25) существует и единственно.

Рассмотрим вопрос об устойчивости разностной схемы (7.25). Используем метод гармоник. Однородное уравнение имеет следующий вид

$$y_j^{n+1} - 2y_j^n + y_j^{n-1} = \frac{\tau^2}{h^2}\left(y_{j+1}^n - 2y_j^n + y_{j-1}^n\right).$$

Подставим в него частное решение вида $y_j^n = q^n e^{i\,jh\varphi}$. Тогда получим:

$$q^2 - 2q + 1 = \frac{\tau^2}{h^2}q(e^{ih\varphi} - 2 + e^{-ih\varphi}) \quad \Longleftrightarrow \quad q^2 - \left(2 - \frac{\tau^2}{h^2}\sin^2\frac{h\varphi}{2}\right)q + 1 = 0.$$

Это уравнение имеет два корня

$$q_{1,2} = \left(1 - 2\frac{\tau^2}{h^2}\sin^2\frac{h\varphi}{2}\right) \pm \sqrt{\left(1 - 2\frac{\tau^2}{h^2}\sin^2\frac{h\varphi}{2}\right)^2 - 1}.$$

Возможны два случая значений $D = \left(1 - 2\frac{\tau^2}{h^2}\sin^2\frac{h\varphi}{2}\right)^2 - 1$:

1. $D > 0$

   Корни уравнения являются вещественными числами, для которых выполнено равенство $q_1 \cdot q_2 = 1$. Следовательно, для одного из корней верно неравенство $|q| > 1$, то есть устойчивости в данном случае нет.

2. $D \leqslant 0$

   Корни уравнения являются комплексными числами, для которых квадрат модуля равен

$$|q_{1,2}|^2 = \left(1 - 2\frac{\tau^2}{h^2}\sin^2\frac{h\varphi}{2}\right)^2 + 1 - \left(1 - 2\frac{\tau^2}{h^2}\sin^2\frac{h\varphi}{2}\right)^2 = 1.$$

   Следовательно, необходимое условие устойчивости разностной схемы выполнено.

Неравенство $D \leqslant 0$ реализуется, если $|1 - 2\frac{\tau^2}{h^2}\sin^2\frac{h\varphi}{2}| \leqslant 1$. Итак, должны выполняться два неравенства $-1 \leqslant 1 - 2\frac{\tau^2}{h^2}\sin^2\frac{h\varphi}{2} \leqslant 1$. Правое неравенство выполнено всегда, а левое неравество приводится к виду $\frac{\tau^2}{h^2}\sin^2\frac{h\varphi}{2} \leqslant 1$. Отсюда следует, что необходимое условие устойчивости выполнено, если $\dfrac{\tau}{h} \leqslant 1$.

Таким образом, разностная схема (7.25) является условно устойчивой и её решение имеет второй порядок точности по τ и h.

7.3 Разностная аппроксимация задачи Дирихле для уравнения Пуассона

Рассмотрим задачу Дирихле для уравнения Пуассона в прямоугольнике $0 < x_1 < l_1,\ 0 < x_2 < l_2$:

$$\begin{cases} \dfrac{\partial^2 u(x_1,x_2)}{\partial x_1^2} + \dfrac{\partial^2 u(x_1,x_2)}{\partial x_2^2} = -f(x_1,x_2), \quad 0 < x_1 < l_1,\ 0 < x_2 < l_2; \\ u(0,x_2) = \mu_1(x_2), \quad u(l_1,x_2) = \mu_2(x_2), \quad 0 < x_2 < l_2; \\ u(x_1,0) = \overline{\mu}_1(x_1), \quad u(x_1,l_2) = \overline{\mu}_2(x_1), \quad 0 < x_1 < l_1. \end{cases} \tag{7.27}$$

Введём в прямоугольной области набор дискретных точек с координатами $(x_{1,i}, x_{2,j})$, где $x_{1,i} = ih_1,\ i = \overline{0,\ N_1},\ h_1 = \dfrac{l_1}{N_1}$ и $x_{2,j} = jh_2,\ j = \overline{0,\ N_2},\ h_2 = \dfrac{l_2}{N_2}$.

Частные производные второго порядка в задаче (7.27) приближённо заменим на вторые разностные производные:

$$\left.\frac{\partial^2 u(x_1,x_2)}{\partial x_1^2}\right|_{x_1=x_{1,i},\,x_2=x_{2,j}} \approx \frac{1}{h_1^2}\left(u(x_{1,i+1},x_{2,j}) - 2u(x_{1,i},x_{2,j}) + u(x_{1,i-1},x_{2,j})\right);$$

$$\left.\frac{\partial^2 u(x_1,x_2)}{\partial x_2^2}\right|_{x_1=x_{1,i},\,x_2=x_{2,j}} \approx \frac{1}{h_2^2}\left(u(x_{1,i},x_{2,j+1}) - 2u(x_{1,i},x_{2,j}) + u(x_{1,i},x_{2,j-1})\right),$$

$$i = \overline{1,\,N_1-1},\ j = \overline{1,\,N_2-1}.$$

Приближённое значение функции $u(x_1,x_2)$ в точке $(x_{1,i},x_{2,j})$ обозначим через y_{ij}.

Сопоставим краевой задаче (7.27) разностную схему вида

$$\begin{cases} y_{\overline{x_1}x_1,ij} + y_{\overline{x_2}x_2,ij} = \varphi_{ij}, \quad i = \overline{1,\,(N_1-1)},\ j = \overline{1,\,(N_2-1)}; \\ y_{0\,j} = \mu_{1,\,j}, \quad y_{N_1\,j} = \mu_{2,\,j}, \quad j = \overline{1,\,(N_2-1)}; \\ y_{i\,0} = \overline{\mu}_{1,\,i}, \quad y_{i\,N_2} = \overline{\mu}_{2,\,i}, \quad i = \overline{1,\,(N_1-1)}. \end{cases} \tag{7.28}$$

Здесь $\mu_{1,\,j} = \mu_1(x_{2,j})$, $\mu_{2,\,j} = \mu_2(x_{2,j})$, $\overline{\mu}_{1,\,i} = \overline{\mu}_1(x_{1,i})$, $\overline{\mu}_{2,\,i} = \overline{\mu}_2(x_{1,i})$ и $\varphi_{ij} = -f(x_{1,i},x_{2,j}) + O(h_1^2+h_2^2)$.

Разностная схема (7.28) построена на 5-ти точечном шаблоне (см. Рис.7.5) и аппроксимирует краевую задачу (7.27) с 2-ым порядком по h_1 и h_2.

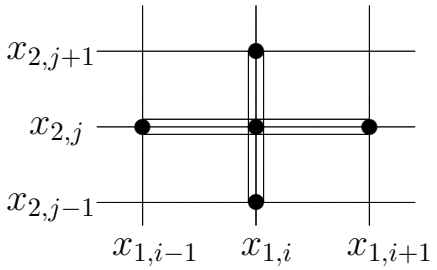

*Рис.*7.5

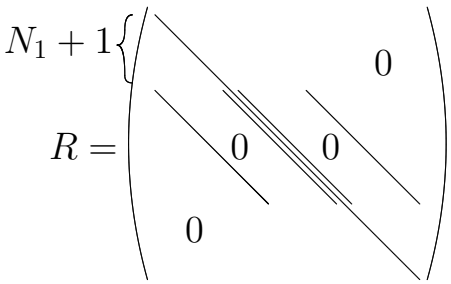

*Рис.*7.6

Введём в рассмотрение вектор y, компонентами которого являются неизвестные y_{ij}, упорядоченные по возрастанию индекса j, и при одинаковом j по возрастающему индексу i, то есть

$$y = \big(y_{10},\, y_{20},\, \ldots,\, y_{(N_1-1)0},\ y_{01},\, y_{11},\, \ldots,\, y_{N_1 1},\ y_{02},\, y_{12},\, \ldots,\, y_{N_1 2},\ \ldots$$
$$\ldots,\, y_{0(N_2-1)},\, y_{1(N_2-1)},\, \ldots,\, y_{N_1(N_2-1)},\ y_{1N_2},\, y_{2N_2},\, \ldots,\, y_{(N_1-1)N_2}\big)^T.$$

Тогда система уравнений (7.28) в матричной форме записи примет вид

$$Ry = g,$$

где вектор правых частей равен

$$g = \big(\overline{\mu}_{1,1},\, \overline{\mu}_{1,2},\, \ldots, \overline{\mu}_{1,(N_1-1)},\; \mu_{1,1},\, \varphi_{11},\, \varphi_{21},\, \ldots, \varphi_{(N_1-1)1},\, \mu_{2,1},$$
$$\mu_{1,2},\, \varphi_{12},\, \varphi_{22},\, \ldots, \varphi_{(N_1-1)2},\, \mu_{2,2},\; \ldots,\; \mu_{1,(N_2-1)},\, \varphi_{1(N_2-1)},\, \varphi_{2(N_2-1)},\, \cdots$$
$$, \varphi_{(N_1-1)(N_2-1)},\, \mu_{2,(N_2-1)},\; \overline{\mu}_{2,1},\, \overline{\mu}_{2,2},\, \ldots,\, \overline{\mu}_{2,(N_1-1)}\big)^T,$$

а квадратная матрица R имеет «ленточную» структуру (см. Рис.7.6), которая соответствует выбранному порядку расположения неизвестных в компонентах вектора y. Размерность матрицы R равна $N \times N$, где $N = N_1 \cdot N_2 + N_1 + N_2 - 3$. Количество компонент вектора неизвестных y также равно N. Если, например, $N_1 = N_2 = 100$ ($l_1 = l_2 = 1 \Rightarrow h_1 = h_2 = 0.01$), то $N > 10\,000$. Эффективными методами решения систем линейных алгебраических уравнений такой размерности являются итерационные методы (см. Глава 1).

Литература

[1] Гантмахер Ф.Р. Теория матриц. — М.: Государственное издательство технико-теоретической литературы, 1953.

[2] Самарский А.А. Теория разностных схем. – 3-е изд. — М.: Наука, 1989.

[3] Самарский А.А., Гулин А.В. Численные методы. — М.: Наука, 1989.

[4] Годунов С.К., Рябенький В.С. Разностные схемы (введение в теорию). — М.: Наука, 1977.

[5] Ильин В.А., Позняк Э.Г. Линейная алгебра: Учебник для вузов. – 4-е изд. — М.: Наука. Физматлит, 1999.

[6] Калиткин Н.Н. Численные методы. — М.: Наука, 1978.

[7] Фаддеев Д.К., Фаддеева В.Н. Вычислительные методы линейной алгебры. — Изд-во «Лань», 2002.

[8] Николаев Е.С. Методы решения сеточных уравнений. — М.: МАКС Пресс, 2018.

[9] Эстербю О., Златев З. Прямые методы для разреженных матриц. — М.: Мир, 1987.

[10] Андреев В.Б. Численные методы: Учебное пособие. — М.: МАКС Пресс, 2013.

[11] Белов А.А., Калиткин Н.Н., Кузьмина Л.В. Сравнение высокоустойчивых форм итерационных методов сопряженных направлений. — Математическое моделирование, 2015, том 27, номер 9, 110-136.

Предметный указатель

Чебышевский набор итерационных параметров, 34
Диагональное преобладание, 10
Элемент наилучшего приближения, 95
Функция ошибок, 52
Интерполирование функций
  кубические сплайны, 87
  полином Эрмита, 85
  полином Лагранжа, 84
Интерполянта, 83
Коэффициент Фурье, 97
Кратные узлы интерполирования, 84
Критерий Сильвестра, 9
Квадратный корень матрицы, 24
Локализация корней нелинейного уравнения, 66
Матрица перехода, 22
Матричная норма вектора, 24
Метод гармоник, 139
Минимальное (максимальное) собственное значение, 63
Модельная задача, 27
Наилучшее приближение в гильбертовом пространстве, 94
Недоопределенная система уравнений, 55
Одношаговый итерационный метод, 15
Определитель Вандермонда, 84
Ортогональная матрица, 58
Переопределенная система уравнений, 56
Погрешность итерационного приближения, 16
Положительно определенная матрица
  достаточное условие, 10
  необходимое условие, 10
Расчет собственных значений матрицы
  метод обратных итераций, 64
  метод вращений, 57
  степенной метод, 61
Разложение Холецкого, 11
Разностная схема для уравнения эллиптического типа, 153
Разностная схема для уравнения колебаний, 151
Разностная схема для уравнения теплопроводности
  нелинейного, 150
  неявная схема, 140
  с переменными коэффициентами, 148
  схема с весами, 143
  явная схема, 133
Разностное уравнение, 101
Решение нелинейных уравнений
  метод бисекции, 67
  метод Ньтона для системы уравнений, 77
  метод Ньютона, 69
  метод простой итерации, 67
  метод секущих, 70
  модифицированный метод Ньтона, 70

Решение систем линейных уравнений
метод Гаусса, 9
метод Крейга, 55
метод квадратного корня, 11
метод минимальных невязок, 44
метод минимальных погрешностей, 45
метод минимальных поправок, 45
метод простой итерации, 19
метод релаксации, 18
метод Ричардсона, 19
метод скорейшего спуска, 44
метод сопряженных градиентов, 46
метод Якоби, 17
метод Зейделя, 18
модифицированный метод квадратного корня, 13
попеременно–треугольный итерационный метод, 30
симметризованные сопряжённые градиенты, 55
Решение задачи Коши для обыкновенного дифференциального уравнения
методы Адамса, 111
методы Гира, 114
методы Рунге-Кутта, 102
Сходимость интерполяционного процесса, 86
Сходимость метода Ньютона, 75
Сходимость метода простой итерации, 72
Сходящийся итерационный метод, 16
Симметричная матрица, 9
Скорость сходимости итерационного метода, 23
Собственные значения матрицы, 57
Стационарный итерационный метод, 16
Шаблон разностной схемы, 134
Шаг дискретной сетки, 100
Упорядоченный набор параметров, 37
Ускорение сходимости (Метод Эйткена), 74
Узел дискретной сетки, 100
Узлы интерполирования, 83
Явный итерационный метод, 16
Явный многошаговый метод, 109
Жесткая система дифференциальных уравнений, 120