

Министерство образования и науки РФ

Федеральное государственное бюджетное образовательное
учреждение высшего профессионального образования
Томский государственный университет систем управления и
радиоэлектроники (ТУСУР)

Кафедра экономической математики, информатики и статистики

И.Г. БОРОВСКОЙ

ЧИСЛЕННЫЕ МЕТОДЫ

Учебно-методическое пособие

Томск-2017

Боровской И.Г.

Численные методы. Учебно-методическое пособие для студентов направления 09.04.02 - Информационные системы и технологии (Управление проектами в информационных системах) / И.Г. Боровской. – Томск: ТУСУР, 2017. – 109 с.

Учебно-методическое пособие разработано в соответствии с решением кафедры экономической математики, информатики и статистики для студентов магистерской подготовки направления 09.04.02 - Информационные системы и технологии (Управление проектами в информационных системах)

Составитель: профессор И.Г. Боровской

Учебно-методическое пособие утверждено на заседании кафедры экономической математики, информатики и статистики 20 октября 2017 г., протокол № 1

© ТУСУР, каф. ЭМИС

© Боровской И.Г.

СОДЕРЖАНИЕ

	Стр
Введение.....	5
1. Численное решение нелинейных уравнений	6
1.1 Постановка задачи. Основные этапы решения	6
1.2 Метод дихотомии и метод простой итерации.....	10
1.3. Метод Ньютона и его модификации.....	12
2. Численное решение систем линейных алгебраических уравнений.....	19
2.1. Норма матрицы. Обусловленность систем линейных алгебраических уравнений.....	19
2.2. Метод Гаусса	23
2.3. Метод прогонки	26
2.4. Метод Якоби.....	27
2.5. Метод Зейделя.....	31
2.6. Двухслойные итерационные методы	35
3. Численное решение систем нелинейных уравнений.....	41
3.1. Метод простой итерации.....	42
3.2. Метод Ньютона для решения системы нелинейных уравнений.....	43
3.3. Модификации метода Ньютона для решения системы нелинейных уравнений	46
3.4. Методы спуска для решения системы нелинейных уравнений	47
4. Приближение функций.....	52
4.1. Интерполирование полиномами. Полином Лагранжа	52
4.2. Интерполяция с кратными узлами. Полином Эрмита	55
4.3. Конечные разности. Интерполяционный полином Ньютона для равноотстоящих узлов	56
4.4. Минимизация оценки погрешности интерполяции. Многочлены Чебышева.....	58
4.5. Интерполяция сплайнами	60
4.6. Метод наименьших квадратов.....	62
5. Численное интегрирование.....	67
5.1. Простейшие квадратурные формулы. Формула прямоугольников, трапеции и формула Симпсона	67
5.2. Оценка погрешности квадратурных формул прямоугольников, трапеции и формулы Симпсона	70
5.3. Квадратурные формулы интерполяционного типа	72
5.4. Квадратурные формулы Гаусса. Полиномы Лежандра	74
5.5. Апостериорные оценки погрешности.....	77
5.6. Построение первообразной с помощью численного интегрирования ..	79

6. Численные методы решения задачи Коши для обыкновенных дифференциальных уравнений.....	81
6.1. Разностная аппроксимация производных	81
6.2. Метод Эйлера	83
6.3. Метод Рунге-Кутты	86
6.4. Метод Адамса.....	90
6.5. Устойчивость численных методов решения задачи Коши.....	92
6.6. Понятие о жестких задачах	96
7. Разностные методы решения уравнения теплопроводности.....	97
7.1. Разностные схемы для одномерного уравнения теплопроводности	98
7.2. Устойчивость разностных схем.....	100
8. МЕТОД КОНЕЧНЫХ ЭЛЕМЕНТОВ.....	103
Библиографический список	108

ВВЕДЕНИЕ

Дисциплина «Численные методы» относится к математическому и естественно-научному циклу (вариативная часть) дисциплин, задачей которой является формирование у студентов теоретических знаний и практических навыков в использовании численных методов при решении задач поиска нулей функций одной переменной, решения систем линейных и нелинейных уравнений, вычисления собственных чисел и собственных векторов матриц, обращения матриц, интерполирования функций, численного дифференцирования и интегрирования функций, решения дифференциальных и интегральных уравнений.

Предметом изучения в рассматриваемой дисциплине являются модели и алгоритмы численного решения прикладных задач из различных областей науки и техники.

Целью дисциплины является изучение теоретических методов и освоение практических навыков в использовании численных методов при решении различных прикладных задач.

Требования к результатам освоения дисциплины

Процесс изучения дисциплины «Численные методы» направлен на формирование компетенции ОПК-6: способность анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями.

В результате освоения содержания дисциплины «Численные методы» студент должен:

Знать

- особенности математических вычислений, реализуемых на ЭВМ;
- теоретические основы численных методов, устойчивость и сложность алгоритма;
- численные методы линейной алгебры;
- решение нелинейных уравнений и систем;
- численное интегрирование и дифференцирование;
- методы приближения функции;
- методы решения дифференциальных уравнений;
- методы решения интегральных уравнений;

Уметь

- строить алгоритмы реализации численных методов решения прикладных программ;
- разрабатывать программы, реализующие численные методы.

Владеть

- навыками применения базового инструментария численных методов для решения прикладных задач;
- методикой построения, анализа и применения численных моделей в профессиональной деятельности.

1. ЧИСЛЕННОЕ РЕШЕНИЕ НЕЛИНЕЙНЫХ УРАВНЕНИЙ

1.1 Постановка задачи. Основные этапы решения

Существует множество классов уравнений — алгебраические, трансцендентные, дифференциальные, интегральные, функциональные, операторные и проч. Конечно, предпочтительными являются аналитические методы решения, позволяющие получить его в виде формулы. Примеры уравнений, позволяющих получать аналитические решения, хорошо известны студентам, например, из курса высшей математики.

Из алгебраических это линейные и квадратные уравнения, из трансцендентных — простейшие показательные, логарифмические и тригонометрические уравнения. Простейшие дифференциальные уравнения — с разделяющимися переменными, линейные, однородные и некоторые другие — также позволяют получать аналитические решения. При изучении других классов уравнений, например, интегральных, удастся выделить некоторые простейшие их разновидности, которые, с известными оговорками, решаются аналитически. Тем не менее подавляющее большинство уравнений, встречающихся в приложениях, не могут быть решены аналитически.

Численные методы решения уравнений являются гораздо более мощными, нежели аналитические. Они тоже не всемогущи, но в умелых руках численные методы позволяют получать решения множества уравнений, совершенно недоступных для аналитических методов. При этом надо заметить, что указанная недоступность может быть обусловлена двумя обстоятельствами: недостаточным уровнем математического образования того, кто решает уравнение, и принципиальной невозможностью решения.

Рассмотрим нелинейное уравнение $f(x) = 0$. Корнем уравнения называется такое число c , что $f(c) = 0$. Корень будет простым, если $f'(c) \neq 0$, и кратным порядка m , если $f^{(k)}(c) = 0$, $k = 1, 2, \dots, m-1$. Графически простой корень соответствует точке пересечения графика функции $y=f(x)$ с осью Ox под ненулевым углом, а кратный корень соответствует точке пересечения под нулевым углом (рис. 1).

Решение задачи отыскания корней нелинейного уравнения осуществляется в два этапа. Первый этап — этап *отделения корней*, второй — этап *итерационного уточнения корней*.

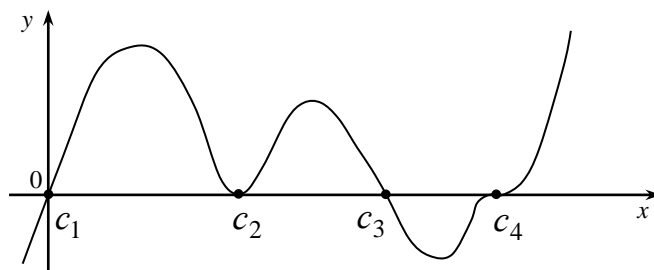


Рис. 1

В большинстве случаев отделение корней можно провести графически. В сомнительных случаях графическое отделение корней необходимо подкрепить вычислениями. При этом используются очевидные положения:

1) если непрерывная функция принимает на концах отрезка $[a, b]$ значения разных знаков (т. е. $f(a) \cdot f(b) < 0$), то уравнение имеет на этом отрезке, по крайней мере, один корень;

2) если же непрерывная функция к тому же строго монотонна, то корень на отрезке единственный.

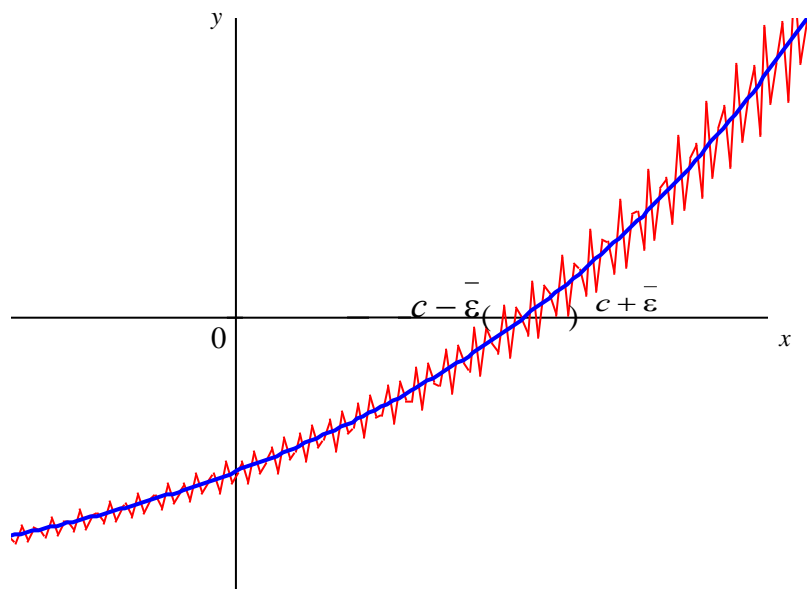


Рис. 2

Будем считать, что входными данными для вычисления корня c являются значения $f(x)$ в окрестности корня. Погрешности в значениях $f^*(x)$ могут быть связаны с неизбежными ошибками округления и с использованием для вычисления $f^*(x)$ приближенных методов.

Пусть в окрестности корня $|f(x) - f^*(x)| < \bar{\Delta}$. Как видно на рис. 2, найдется окрестность корня c , $(c - \bar{\varepsilon}, c + \bar{\varepsilon})$, в которой выполняется неравенство $|f(x)| < \bar{\Delta}$. Будем называть этот интервал радиуса $\bar{\varepsilon}$ *интервалом неопределенности корня c* . Оценим величину $\bar{\varepsilon}$.

Из приближенного равенства $f(x) \approx f(c) + f'(c)(x - c) = f'(c)(x - c)$ следует $|f'(c)(x - c)| < \bar{\Delta}$, откуда получаем

$$c - \frac{\bar{\Delta}}{|f'(c)|} < x < c + \frac{\bar{\Delta}}{|f'(c)|}.$$

Следовательно, $\bar{\varepsilon} = \frac{\bar{\Delta}}{|f'(c)|}$ и $\nu_{\Delta} = \frac{1}{|f'(c)|}$ – абсолютное число обусловленности.

Для корня кратности m в силу формулы Тейлора справедливо приближенное равенство

$$f(x) \approx \frac{f^{(m)}(c)}{m!} (x - c)^m.$$

Следовательно, $\bar{\varepsilon} = \left(\frac{\bar{\Delta} m!}{|f^{(m)}(c)|} \right)^{1/m}$, что свидетельствует о плохой обусловленности задачи вычисления кратных корней уравнения.

В условиях неопределенности, вызванных приближенным заданием функции, любое значение из интервала неопределенности корня может быть принято за решение уравнения.

1.2 Метод дихотомии и метод простой итерации

Пусть уравнение $f(x) = 0$ на отрезке $[a, b]$ имеет единственное решение, причем $f(x)$ непрерывная функция на данном отрезке. Разделим отрезок $[a, b]$ пополам в точке $d = (a+b)/2$. Если $f(d) \neq 0$, то возможны два случая:

- 1) функция $f(x)$ меняет знак на отрезке $[a, d]$;
- 2) функция $f(x)$ меняет знак на отрезке $[d, b]$.

Выбирая в каждом случае отрезок $[a_n, b_n]$, на котором функция $f(x)$ меняет знак, и, продолжая процесс половинного деления дальше, можно прийти до сколь угодно малого отрезка, содержащего корень уравнения.

Середина n -го отрезка – точка $x_n = (a_n + b_n)/2$ дает приближение корню c

$$|x_n - c| \leq (b_n - a_n)/2 = (b - a)/2^{n+1}.$$

Критерием окончания итерационного процесса служит неравенство

$$(b_n - a_n)/2 < \varepsilon.$$

Можно принять $x_n = (a_n + b_n)/2$ за приближение к корню с точностью ε .

Пример 1. Вычислить корень уравнения $x^2 = \cos x$ методом бисекций с точностью $\varepsilon = 0,00001$.

Так как функция $f(x) = x^2 - \cos x$ – четная, и $x = 0$ не является корнем уравнения, то исходное уравнение имеет четное число симметричных корней.

Пусть $a = 0, b = 5$ ($5^2 - \cos 5 > 0, 0^2 - \cos 0 < 0$), тогда $d = (a + b)/2 = 2,5$. Следующий отрезок, на котором функция меняет знак, есть $[a_1, b_1] = [0; 2,5]$. Делим отрезок пополам, при этом $d = 1,25$, и получаем следующий отрезок $[a_2, b_2] = [0; 1,25]$. Делим отрезок $[a_2, b_2] = [0; 1,25]$ пополам, при этом $d = 0,625$ и следующий отрезок, на котором функция меняет знак, есть $[a_3, b_3] = [0,625; 1,25]$ и т. д.

Для отрезка $[a_{18}, b_{18}] = [0,824127; 0,824146]$ выполнено условие $(b_n - a_n)/2 < \varepsilon$, поэтому можно принять $x_{18} = (a_{18} + b_{18})/2 = 0,82414$ за приближение к корню с заданной точностью ε .

Чтобы применить метод простой итерации необходимо преобразовать уравнение $f(x) = 0$ к следующему виду:

$$x = \varphi(x). \quad (1.1)$$

Функцию $\varphi(x)$ далее будем называть итерационной функцией. Выберем приближенное значение корня x_0 и будем строить последовательность чисел $\{x_n\}$ (итерационную последовательность), определенную с помощью рекуррентной формулы

$$x_{n+1} = \varphi(x_n). \quad (1.2)$$

Рис. 3, 4 иллюстрируют случаи, когда итерационная последовательность сходится при произвольном начальном приближении. Напротив, рис. 5, 6 показывают, что итерационная последовательность расходится при любом начальном приближении.

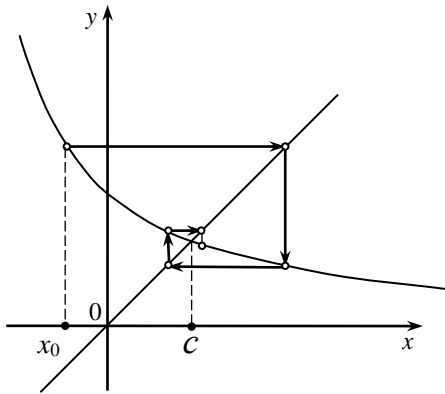


Рис. 3

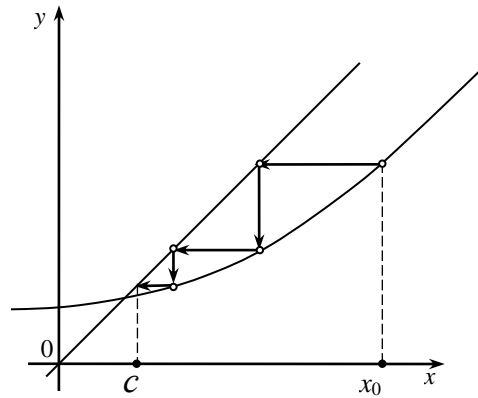


Рис. 4

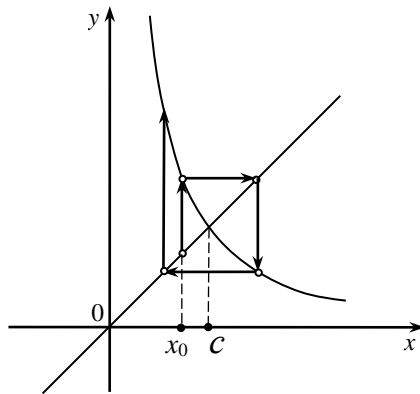


Рис. 5

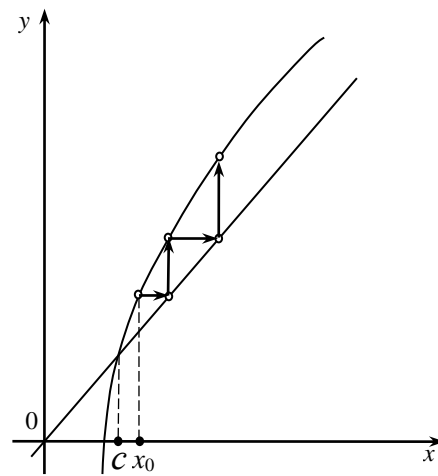


Рис. 6

Теорема о сходимости итерационной последовательности.

Пусть c – корень уравнения (1.1) и пусть функция $\varphi(x)$ удовлетворяет в некоторой σ -окрестности корня c условию Липшица с константой L , $0 < L < 1$

$$|\varphi(x_2) - \varphi(x_1)| \leq L|x_2 - x_1| \quad \forall x_2, x_1 \in [c - \sigma, c + \sigma].$$

Тогда независимо от выбора начального приближения x_0 из указанной σ -окрестности корня итерационная последовательность не выходит из этой окрестности, сходится к корню уравнения (1.1) и имеет место следующая оценка погрешности:

$$|x_n - c| \leq L^n |x_0 - c|. \quad (1.3)$$

Доказательство. Пусть $x_1 = \varphi(x_0)$. При этом будем иметь

$$|x_1 - c| \leq |\varphi(x_0) - \varphi(c)| \leq L |x_0 - c| \leq L\sigma.$$

Продолжим построение итерационной последовательности. Вычислим $x_2 = \varphi(x_1)$.

При этом

$$|x_2 - c| \leq |\varphi(x_1) - \varphi(c)| \leq L |x_1 - c| \leq L^2\sigma.$$

По индукции легко показать, что все последующие итерации не выводят последовательность $\{x_n\}$ из σ -окрестности корня c и удовлетворяют неравенствам

$$|x_n - c| \leq L^n |x_0 - c| \leq L^n \sigma.$$

Откуда следует, что

$$\lim_{n \rightarrow \infty} x_n = c.$$

Замечание 1. Условие Липшица с константой $L < 1$ будет заведомо обеспечиваться, если предположить, что $\varphi(x)$ на $(c - \sigma, c + \sigma)$ имеет непрерывную производную и $|\varphi'(x)| \leq q < 1$. В этом случае $L = q$.

Выведем *апостериорную* оценку погрешности, пригодную для практического применения в качестве критерия окончания.

Пусть $|\varphi'(x)| \leq q < 1$. Из формулы конечных приращений Лагранжа, получаем

$$x_n - c = \varphi(x_{n-1}) - \varphi(c) = \varphi'(\xi_n)(x_{n-1} - c) = \varphi'(\xi_n)(x_{n-1} - x_n) + \varphi'(\xi_n)(x_n - c),$$

где ξ_n находится между c и x_{n-1} .

Откуда

$$|x_n - c| = \left| \frac{\varphi'(\xi_n)}{1 - \varphi'(\xi_n)} \right| |x_n - x_{n-1}| \leq \frac{q}{1 - q} |x_n - x_{n-1}|.$$

Если значение q известно, то вычисление корня с точностью ε следует вести до выполнения неравенства $\frac{q}{1 - q} |x_n - x_{n-1}| < \varepsilon$ или равносильного ему неравенства

$$|x_n - x_{n-1}| < \frac{1 - q}{q} \varepsilon. \quad (1.4)$$

Чтобы применить метод простой итерации, необходимо преобразовать уравнение $f(x) = 0$ к виду $x = \varphi(x)$.

Для этого положим

$$x = x - mf(x), \text{ или } \varphi(x) = x - mf(x).$$

Для выполнения условий теоремы 1.1, достаточно подобрать m так, чтобы в σ -окрестности корня c выполнялось неравенство $|\varphi'(x)| = |1 - mf'(x)| < 1$.

1.3. Метод Ньютона и его модификации

Выведем расчетную формулу метода для решения нелинейного уравнения $f(x) = 0$ из простых геометрических соображений. Иллюстрация метода приведена на рис. 7.

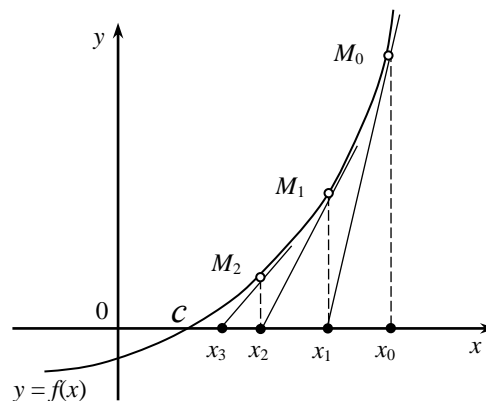


Рис. 7

Пусть x_0 — заданное начальное приближение к корню c . В точке $M_0(x_0, f(x_0))$ проведем касательную к графику функции $y = f(x)$ и за новое приближение x_1 примем абсциссу точки пересечения этой касательной с осью Ox . За приближение x_2 примем абсциссу точки пересечения с осью Ox касательной, проведенной к графику функции в точке $M_1(x_1, f(x_1))$. Продолжая этот процесс далее, получим последовательность $\{x_n\}$ приближений метода Ньютона к корню c .

Уравнение касательной к графику функции $y = f(x)$ в точке $M_n(x_n, f(x_n))$ имеет вид

$$y = f(x_n) + f'(x_n)(x - x_n).$$

Полагая в уравнении касательной $y = 0$, получим расчетную формулу для x_{n+1} метода Ньютона:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad n \geq 0. \quad (1.5)$$

Теорема о сходимости метода Ньютона. Пусть c – простой корень уравнения $f(x) = 0$, в некоторой окрестности которого функция дважды непрерывно дифференцируема.

Тогда найдется такая малая σ -окрестность корня c , что при произвольном выборе начального приближения x_0 из этой окрестности итерационная последовательность метода Ньютона не выходит за пределы этой окрестности и сходится к корню c .

Доказательство. Пусть $|f'(x)| \geq m > 0$, $|f''(x)| \leq M$, $x \in [a, b]$. Уравнение $f(x) = 0$ на отрезке $[a, b]$ эквивалентно уравнению $x = \varphi(x)$, где $\varphi(x) = x - \frac{f(x)}{f'(x)}$. Вычислим и оценим производную функции $\varphi(x)$:

$$\begin{aligned} \varphi'(x) &= 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2}, \\ |\varphi'(x)| &\leq \frac{|f(x)|M}{m^2}. \end{aligned}$$

Поскольку функция $f(x)$ непрерывна и $f(c) = 0$, то найдется малая σ -окрестность корня c , что будет выполняться неравенство

$$|f(x) - f(c)| = |f(x)| \leq \varepsilon = \frac{m^2}{2M}.$$

Учитывая это, получаем окончательную оценку производной:

$$|\varphi'(x)| \leq 1/2, \quad x \in [c - \sigma, c + \sigma].$$

Это означает, что выполнены условия теоремы 1.1 в малой σ -окрестности корня c .

Выведем *апостериорную* оценку погрешности, пригодную для практического применения в качестве критерия окончания метода Ньютона.

Поскольку $|\varphi'(x)| \leq q = 1/2$, то $\frac{1-q}{q} > 1$ и из оценки (1.4) для метода простой итерации следует, что в качестве критерия окончания метода Ньютона можно использовать простую оценку

$$|x_n - x_{n-1}| < \varepsilon. \quad (1.6)$$

Замечание 2. Практическое применение метода Ньютона имеет две существенные трудности. Одна из них состоит в необходимости вычисления производной $f'(x)$. Вторая трудность состоит в том, что для сходимости необходимо выбирать хорошее начальное приближение, попадающее в малую σ -окрестность корня.

Неудачный выбор начального приближения может дать расходящуюся последовательность (рис. 8).

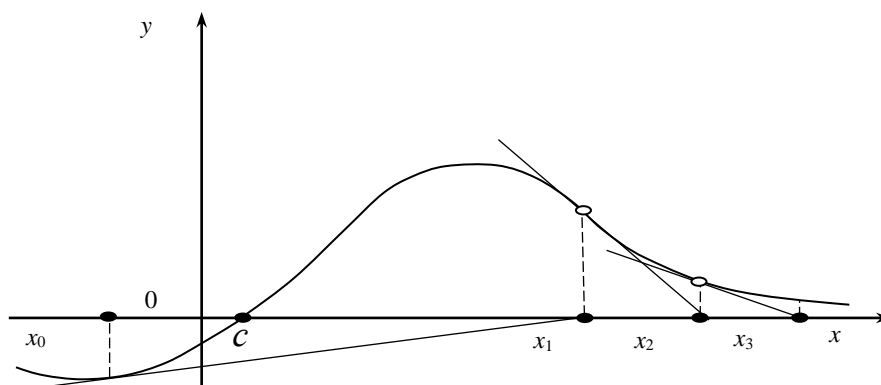


Рис. 8

Замечание 3. Метод Ньютона сходится с квадратичной скоростью $|x_{n+1} - c| \leq C|x_n - c|^2$.

Пример 2. На отрезке $[0, 1]$ найти приближенное значение уравнения $f(x) = x - \cos x = 0$ методом Ньютона.

Рекуррентная формула (1.5) в данном случае принимает вид

$$x_{n+1} = x_n - \frac{x_n - \cos x_n}{1 + \sin x_n}.$$

Выберем в качестве первого приближения $x_0 = 0,5$.

Тогда

$$x_1 = 0,7552224171 \text{ 06};$$

$$x_2 = 0,7391416661 \text{ 50};$$

$$x_3 = 0,7390851339 \text{ 21};$$

$$x_4 = 0,7390851332 \text{ 15};$$

$$x_5 = 0,7390851332 \text{ 15}.$$

Приведенный пример показывает очень высокую скорость сходимости метода Ньютона.

Пример 3. Вычислить $\sqrt{2}$ как задачу решения методом Ньютона уравнения $f(x) = x^2 - 2 = 0$.

Рекуррентная формула (1.5) в данном случае принимает вид

$$x_{n+1} = x_n - \frac{x_n^2 - 2}{2x_n}.$$

Выберем в качестве первого приближения $x_0 = 2$.

Тогда

$$x_1 = 1,5;$$

$$x_2 = 0,416666;$$

$$x_3 = 0,414216.$$

Третья итерация определяет $\sqrt{2}$ с абсолютной погрешностью $\Delta = \sqrt{2} - x_3 = -0,000002$.

Замена в формуле (1.5) метода Ньютона производной приближением $f'(x_n) \approx \frac{f(x_{n-1}) - f(x_n)}{x_{n-1} - x_n}$ приводит к расчетной формуле *метода секущих*:

$$x_{n+1} = x_n - \frac{x_{n-1} - x_n}{f(x_{n-1}) - f(x_n)} f(x_n), \quad n \geq 0. \quad (1.7)$$

Пример 4. Вычислить корень уравнения $x^2 = \cos x$ методом секущих с точностью $\varepsilon = 0,00001$.

Так как $f(x) = x^2 - \cos x$, то формула (1.7) примет вид

$$x_{n+1} = x_n - \frac{x_{n-1} - x_n}{(x_{n-1}^2 - \cos x_{n-1}) - (x_n^2 - \cos x_n)} (x_n^2 - \cos x_n).$$

Положим $x_0 = 5$, $x_1 = 10$, тогда по формуле метода секущих последовательно вычисляем

$$x_2 = 3,376547 ;$$

$$x_3 = 2,450130 ;$$

.....

$$x_8 = 0,824190 ;$$

$$x_9 = 0,824132 ;$$

$$x_{10} = 0,824132 .$$

Критерий окончания итераций $|x_{10} - x_9| < \varepsilon$ выполнен, поэтому можно принять $x_{10} = 0,82413$ за приближение к корню с заданной точностью ε .

Метод секущих является двухшаговым (необходимо задать две начальные точки x_0 и x_1 из окрестности решения уравнения) и обладает локальной сходимостью. Геометрическая иллюстрация метода секущих приведена на рис. 9.

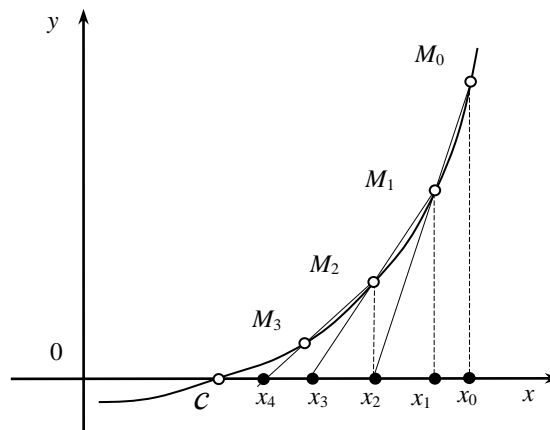


Рис. 9

Замена в формуле (1.5) метода Ньютона производной приближением $f'(x_n) \approx \frac{f(z_n) - f(x_n)}{z_n - x_n}$, где $z_n = x_n + f(x_n)$, приводит к расчетной формуле метода Стеффенсена:

$$x_{n+1} = x_n - \frac{f(x_n)}{f(x_n + f(x_n)) - f(x_n)} f(x_n), \quad n \geq 0. \quad (1.8)$$

Пример 5. Вычислить корень уравнения $x^3 = \cos(x/2)$ методом Стеффенсена с точностью $\varepsilon = 0,00001$.

Так как

$$f(x) = x^3 - \cos(x/2),$$

то формула (1.8) примет вид

$$x_{n+1} = x_n - \frac{(f(x_n))^2}{(x_n + f(x_n))^3 - \cos((x_n + f(x_n))/2) - (f(x_n))},$$

$$f(x_n) = x_n^3 - \cos(x_n/2).$$

Пусть $x_0 = 3$, тогда

$$x_1 = 2,972924 ,$$

$$x_2 = 2,945241 ,$$

$$x_3 = 2,916919 ,$$

.....

$$x_{38} = 0,961912 ,$$

$$x_{39} = 0,960759 ,$$

$$x_{40} = 0,960753 .$$

Критерий окончания итераций $|x_{40} - x_{39}| < \varepsilon$ выполнен, поэтому можно принять $x_{40} = 0,96075$ за приближение к корню с заданной точностью ε .

Для вычисления *корня уравнения кратности m* используют следующую модификацию метода Ньютона, сохраняющую квадратичную скорость сходимости:

$$x_{n+1} = x_n - m \frac{f(x_n)}{f'(x_n)}, \quad n \geq 0. \quad (1.9)$$

Замена в формуле (1.5) метода Ньютона производной приближением $f'(x_n) \approx \frac{f(d) - f(x_n)}{d - x_n}$, где d – фиксированная точка из окрестности простого корня, приводит к расчетной формуле *метода ложного положения*:

$$x_{n+1} = x_n - \frac{d - x_n}{f(d) - f(x_n)} f(x_n), \quad n \geq 0. \quad (1.10)$$

Метод ложного положения имеет линейную сходимость $|x_{n+1} - c| \leq C|x_n - c|$.

Геометрическая иллюстрация метода приведена на рис. 10.

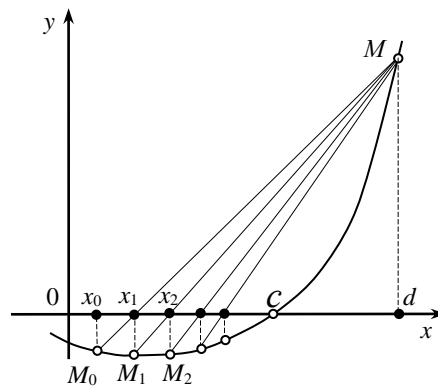


Рис. 10

Пример 6. Вычислить корень уравнения $x^2 = \cos x$ методом ложного положения с точностью $\varepsilon = 0,00001$. Для функции $f(x) = x^2 - \cos x$ в качестве фиксированной точки из окрестности корня выберем $d = 4$, тогда $f(d) = 16,6536$ и формула (1.10) примет вид

$$x_{n+1} = x_n - \frac{4 - x_n}{16,6536 - (x_n^2 - \cos x_n)} (x_n^2 - \cos x_n).$$

Положим $x_0 = 5$, тогда по формуле метода ложного положения последовательно вычисляем

$$x_1 = 1,9934482;$$

$$x_2 = 1,260336;$$

$$x_3 = 1,031664;$$

.....

$$x_{16} = 0,824201;$$

$$x_{17} = 0,824170;$$

$$x_{18} = 0,824153.$$

Критерий окончания итераций $|x_{18} - x_{17}| < \varepsilon$ выполнен, поэтому можно принять $x_{18} = 0,82415$ за приближение к корню с заданной точностью ε . Для вычисления корня с точностью $\varepsilon = 0,00001$ методом ложного положения потребовалось на 10 итераций больше, чем при вычислении методом секущих.

2. ЧИСЛЕННОЕ РЕШЕНИЕ СИСТЕМ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

2.1. Норма матрицы. Обусловленность систем линейных алгебраических уравнений

Рассмотрим задачу определения решения системы линейных алгебраических уравнений (СЛАУ), в которой число неизвестных равно числу уравнений:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n &= b_2, \\ \dots \dots \dots & \dots \dots \dots \\ a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n &= b_n. \end{aligned} \tag{2.1}$$

В матричной форме записи эта система принимает вид

$$A \mathbf{x} = \mathbf{b}$$

где A – невырожденная матрица коэффициентов
– вектор столбец свободных членов

В приложениях часто встречаются матрицы, в которых число ненулевых элементов много меньше общего числа элементов. Такие матрицы называются *разреженными*. Напротив, матрицы общего вида называют *плотными*. В частности, к разреженным относятся диагональные, трехдиагональные матрицы.

Говорят, что в R^n задана норма, если каждому вектору \mathbf{x} из R^n поставлено вещественное число $\|\mathbf{x}\|$, называемое нормой вектора \mathbf{x} и обладающая следующими свойствами:

- 1) $\|\mathbf{x}\| \geq 0$, причем $\|\mathbf{x}\| = 0$, тогда и только тогда, когда $\mathbf{x} = \mathbf{0}$;
- 2) $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$, для любого вектора \mathbf{x} и любого числа α ;
- 3) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$, для любых векторов \mathbf{x} и \mathbf{y} .

В вычислительных методах наиболее употребительными являются следующие нормы:

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{i=n} |x_i|, \quad \|\mathbf{x}\|_2 = \left(\sum_{i=1}^{i=n} x_i^2 \right)^{1/2}, \quad \|\mathbf{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|. \tag{2.3}$$

Указанные нормы являются частными случаями более общей нормы:

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{i=n} |x_i|^p \right)^{1/p},$$

соответственно при $p = 1$, $p = 2$, $p = \infty$.

Замечание 1. Справедливы неравенства

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq m \|\mathbf{x}\|_\infty,$$

указывающие на то, что все три нормы эквивалентны. Норму $\|\mathbf{x}\|_2$ называют евклидовой ($\|\mathbf{x}\|_2 = \|\mathbf{x}\|_E$).

Введем абсолютную и относительную погрешность вектора \mathbf{x}^* с помощью формул

$$\Delta(\mathbf{x}^*) = \|\mathbf{x} - \mathbf{x}^*\|, \quad \delta(\mathbf{x}^*) = \frac{\|\mathbf{x} - \mathbf{x}^*\|}{\|\mathbf{x}\|}. \quad (2.4)$$

Введем величину

$$\|A\| = \max_{\mathbf{x} \neq 0} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}, \quad \|A\| = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|, \quad (2.5)$$

которую принято называть нормой матрицы A , подчиненной норме вектора $\|\cdot\|$.

Известно, что нормам $\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_2$, $\|\mathbf{x}\|_\infty$ подчинены нормы $\|A\|_1$, $\|A\|_2$, $\|A\|_\infty$, вычисляемые по формулам:

$$\begin{aligned} \|A\|_1 &= \max_{1 \leq j \leq n} \sum_{i=1}^{i=n} |a_{ij}|, & \|A\|_2 &= \max_{1 \leq j \leq n} \sqrt{\lambda_j(A^T A)}, \\ \|A\|_\infty &= \max_{1 \leq i \leq n} \sum_{j=1}^{j=n} |a_{ij}|, \end{aligned} \quad (2.6)$$

где $\lambda_j(A^T A)$ – собственные числа матрицы $A^T A$.

Замечание 2. Для оценки нормы $\|A\|_2$ можно использовать неравенство

$$\|A\|_2 = \max_{1 \leq j \leq n} \sqrt{\lambda_j(A^T A)} \leq \sqrt{\sum_{j,i=1}^n a_{ij}^2} = \|A\|_E.$$

Норма $\|A\|$ имеет простую геометрическую интерпретацию. Если рассматривать матрицу A , как матрицу линейного преобразования, которое переводит вектор \mathbf{x} в новый вектор $\mathbf{y} = A\mathbf{x}$, то норма матрицы есть максимальный коэффициент растяжения вектора \mathbf{x} под действием матрицы A . Из формулы (2.5) следует

$$\|A\mathbf{x}\| \leq \|A\| \|\mathbf{x}\|. \quad (2.7)$$

Пусть \mathbf{x}^* решение уравнения $A\mathbf{x}^* = \mathbf{b}^*$, тогда

$$A(\mathbf{x} - \mathbf{x}^*) = \mathbf{b} - \mathbf{b}^* = \mathbf{b} - A(\mathbf{x}^*). \quad (2.8)$$

Невязка $\mathbf{r} = \mathbf{b} - A(\mathbf{x}^*)$ показывает, насколько правая часть системы отличается от левой, если в нее подставить приближенное решение \mathbf{x}^* .

Из формул (2.8), (2.7) находим оценку абсолютной погрешности решения через невязку:

$$\Delta(\mathbf{x}^*) = \|\mathbf{x} - \mathbf{x}^*\| \leq \|A^{-1}\| \|\mathbf{r}\|. \quad (2.9)$$

Замечая, что $\Delta(\mathbf{b}^*) = \|\mathbf{r}\|$, из (2.9) имеем

$$\frac{\Delta(\mathbf{x}^*)}{\|\mathbf{x}\|} \leq \frac{\|A^{-1}\|}{\|\mathbf{x}\|} \Delta(\mathbf{b}^*) = \frac{\|A^{-1}\| \|\mathbf{b}\|}{\|\mathbf{x}\| \|\mathbf{b}\|} \Delta(\mathbf{b}^*).$$

Учитывая, что $\|\mathbf{b}\| = \|A\mathbf{x}\| \leq \|A\| \|\mathbf{x}\|$ и (2.4), получаем оценку для относительной погрешности решения СЛАУ:

$$\delta(\mathbf{x}^*) \leq \|A\| \cdot \|A^{-1}\| \delta(\mathbf{b}^*). \quad (2.10)$$

Число $\text{cond}(A) = \|A\| \cdot \|A^{-1}\|$ называется *числом обусловленности* матрицы A . Чем больше $\text{cond}(A)$, тем резче решение системы реагирует на возмущение правой части. Поэтому матрицы с большим числом обусловленности и соответствующие им СЛАУ называют *плохо обусловленными*.

Расчет нормы матрицы непосредственно по формуле (2.5) достаточно сложен. Выведем оценку числа обусловленности через собственные характеристические числа матрицы. Пусть \mathbf{y} , \mathbf{z} – собственные вектора матрицы A , соответствующие максимальному λ_{\max} и минимальному λ_{\min} по модулю характеристическому числу матрицы A .

Тогда

$$A\mathbf{y} = \lambda_{\max} \mathbf{y}, \quad A\mathbf{z} = \lambda_{\min} \mathbf{z}, \quad \text{или} \quad A^{-1}\mathbf{z} = \mathbf{z} / \lambda_{\min};$$

$$|\lambda_{\max}| \|\mathbf{y}\|_2 = \|A\mathbf{y}\|_2 \leq \|A\|_2 \cdot \|\mathbf{y}\|_2, \quad \|A^{-1}\mathbf{z}\|_2 = \|\mathbf{z}\|_2 / |\lambda_{\min}| \leq \|A^{-1}\|_2 \|\mathbf{z}\|_2.$$

$$\text{Откуда} \quad \|A\|_2 \geq |\lambda_{\max}|, \quad \|A^{-1}\|_2 \geq 1/|\lambda_{\min}|,$$

$$\text{и} \quad \text{cond}(A) = \|A\|_2 \cdot \|A^{-1}\|_2 \geq |\lambda_{\max}| / |\lambda_{\min}|. \quad (2.11)$$

Замечания:

3. Если матрица A симметричная, то все ее характеристические значения вещественны, причем $\|A\| = |\lambda_{\max}|$, $\|A^{-1}\|_2 = 1/|\lambda_{\min}|$ и для таких матриц $\text{cond}(A) = |\lambda_{\max}| / |\lambda_{\min}|$.

4. Если матрица A симметричная и положительная ($(A\mathbf{x}, \mathbf{x}) > 0$, при всех $\mathbf{x} \neq 0$), то все ее характеристические значения вещественны и положительны, $\|A\| = \lambda_{\max}$, $\|A^{-1}\|_2 = 1/\lambda_{\min}$ и для таких матриц $\text{cond}(A) = \lambda_{\max} / \lambda_{\min}$.

5. Число обусловленности тем больше, чем больше разброс характеристических чисел матрицы A , поэтому с увеличением размера матрицы, ее число обусловленности имеет тенденцию к ухудшению.

Пример 1. Рассмотрим систему уравнений
$$\begin{cases} 2x_1 + 4x_2 = 6, \\ 4x_1 + 9x_2 = 13. \end{cases}$$

Легко проверить, что ее решение $x_1 = 1, x_2 = 1$.

Матрица $A = \begin{bmatrix} 2 & 4 \\ 4 & 9 \end{bmatrix}$ симметричная и положительная, т. к. ее главные

миноры положительные (критерий Сильвестра). Ее характеристические значения $\lambda_1 \approx 10,81507$ и $\lambda_2 \approx 0,184927$; $\text{cond}(A) = \lambda_{\max} / \lambda_{\min} = 58,482915$, что говорит о плохой обусловленности матрицы A .

Изменим правую часть рассматриваемой системы
$$\begin{cases} 2x_1 + 4x_2 = 6, \\ 4x_1 + 9x_2 = 12,9. \end{cases}$$

Тогда

$$x_1^* = 1,2, \quad x_2^* = 0,9; \quad \delta(\mathbf{x}^*) = \frac{\|\mathbf{x} - \mathbf{x}^*\|}{\|\mathbf{x}\|} = \frac{\sqrt{0,2^2 + 0,1^2}}{\sqrt{2}} \approx 0,158114.$$

С геометрической точки зрения решение системы есть точка пересечения двух прямых с близкими угловыми коэффициентами, поэтому даже незначительная погрешность в задании положения этих прямых существенно меняет положение точки пересечения (рис. 11).

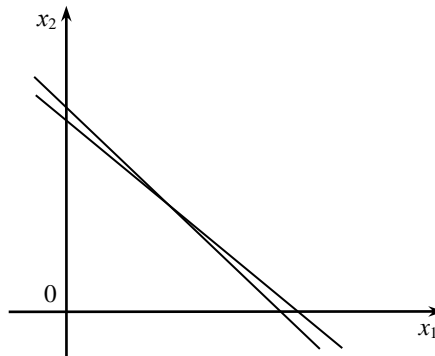


Рис. 11

2.2. Метод Гаусса

Метод Гаусса относится к прямым методам, позволяющим получить решение системы после выполнения конечного числа операций. Метод Гаусса состоит из двух основных этапов, называемых *прямым ходом* и *обратным ходом*. На первом этапе система приводится к треугольному виду. Затем на втором этапе осуществляется последовательное отыскание неизвестных.

Прямой ход состоит из $n - 1$ шага исключения. На первом шаге во всех уравнениях системы (2.1), кроме первого, исключается переменная x_1 .

$$\begin{aligned} x_1 + c_{12}x_2 + \dots + \tilde{n}_{1n}x_n &= y_1, \\ a_{22}^1x_2 + \dots + a_{2n}^1x_n &= b_2^1, \\ &\dots\dots\dots \\ a_{n2}^1x_2 + \dots + a_{nn}^1x_n &= b_n^1. \end{aligned}$$

Здесь

$$c_{12} = \frac{a_{12}}{a_{11}}, \dots, c_{1n} = \frac{a_{1n}}{a_{11}}, \quad y_1 = \frac{b_1}{a_{11}}; \quad (2.12)$$

$$a_{ij}^1 = a_{ij} - a_{i1} \frac{a_{1j}}{a_{11}}, \quad b_i^1 = b_i - a_{i1} \frac{b_1}{a_{11}}.$$

Выделяем укороченную систему

$$a_{22}^1 x_2 + a_{23}^1 x_3 + \dots + a_{2n}^1 x_n = b_2^1,$$

.....

$$a_{n2}^1 x_2 + a_{n3}^1 x_3 + \dots + a_{nn}^1 x_n = b_n^1.$$

Продолжая далее процесс исключения, после $n - 1$ шага редуцируем систему к треугольному виду с верхней треугольной матрицей:

$$\begin{aligned} x_1 + c_{12}x_2 + c_{13}x_3 + \dots + c_{1n}x_n &= y_1, \\ x_2 + c_{23}x_3 + \dots + c_{2n}x_n &= y_2, \\ \dots & \\ x_n &= y_n, \end{aligned} \quad C = \begin{bmatrix} 1 & c_{12} & c_{13} & \dots & c_{1n} \\ 0 & 1 & c_{23} & \dots & c_{2n} \\ 0 & 0 & 1 & \dots & c_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 1 & \dots \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \quad (2.13)$$

Обратный ход состоит в последовательном определении неизвестных из системы (2.13) в обратном порядке.

Описанная выше процедура решения системы может оказаться неустойчивой по отношению к случайным ошибкам. Чтобы избежать этого, естественно потребовать выполнения условия

$$|c_{ij}| < 1. \quad (2.14)$$

Приступая к первому шагу прямого хода, найдем в первой строке матрицы максимальный по модулю коэффициент $|a_{1j}|$. Переставляя первый столбец с j -м, сделаем ведущим элементом a_{11} первого шага наибольший по модулю элемент первой строки. Благодаря этому элементы c_{1j} , вычисленные по (2.12), будут удовлетворять неравенству (2.14).

Продолжая эту процедуру на каждом шаге прямого хода для укороченных систем, обеспечим выполнение неравенства (2.14) для всех элементов треугольной матрицы C . Такой способ коррекции называется *выбором ведущего элемента по строке*.

Поскольку определитель приведенной матрицы C равен 1, то определитель исходной системы

$$\Delta = \det A = (-1)^k a_{11} a_{22} \dots a_{nn}^{n-1}, \quad (2.15)$$

где k – число перестановок столбцов в процессе редукции матрицы A к треугольной матрице C .

Из курса алгебры известно, что система (2.1) имеет единственное решение, тогда и только тогда, когда $\det A \neq 0$. Задача вычисления определителя матрицы в общем случае плохо обусловлена.

Матрицы, обладающие свойством

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}|, \quad (2.16)$$

называются матрицами с *диагональным преобладанием*.

Лемма 1. Система с диагональным преобладанием всегда имеет, и притом единственное решение.

Доказательство. Рассмотрим однородную систему $A\mathbf{x} = \mathbf{0}$.

Предположим, что она имеет нетривиальное решение и $x_k \neq 0$, $|x_k| \geq |x_j|$, $1 \leq j \leq n$. Перепишем k -е уравнение в виде

$$a_{kk} x_k = - \sum_{\substack{j=1 \\ j \neq k}}^n a_{kj} x_j.$$

Откуда получим

$$|a_{kk}| |x_k| \leq \sum_{\substack{j=1 \\ j \neq k}}^n |a_{kj}| |x_j| \leq |x_k| \sum_{\substack{j=1 \\ j \neq k}}^n |a_{kj}|.$$

Сокращая последнее неравенство на множитель $|x_k|$, приходим к противоречию с условием (2.16).

2.3. Метод прогонки

Системы с трехдиагональными матрицами возникают при решении задач математической физики, интерполяции сплайнами и других вычислительных задачах. Метод прогонки относится к прямым методам и является эффективным методом решения СЛАУ с трехдиагональными матрицами вида

$$\begin{pmatrix} b_1 & c_1 & 0 & 0 & 0 & 0 \\ a_2 & b_2 & c_2 & 0 & 0 & 0 \\ 0 & a_3 & b_3 & c_3 & 0 & 0 \\ 0 & 0 & \dots & \dots & \dots & 0 \\ 0 & 0 & 0 & a_{n-1} & b_{n-1} & c_{n-1} \\ 0 & 0 & 0 & 0 & a_n & b_n \end{pmatrix},$$

$$a_i x_{i-1} + b_i x_i + c_i x_{i+1} = d_i, \quad 1 \leq i \leq n, \quad a_1 = 0, \quad c_n = 0. \quad (2.17)$$

Преобразуем первое уравнение при $i = 1$ к виду

$$x_1 = \alpha_1 x_2 + \beta_1,$$

где

$$\alpha_1 = -c_1/b_1, \quad \beta_1 = d_1/b_1. \quad (2.18)$$

Подставим полученное для x_1 выражение во второе уравнение системы:

$$a_2(\alpha_1 x_2 + \beta_1) + b_2 x_2 + c_2 x_3 = d_2.$$

Преобразуем это уравнение к виду

$$x_2 = \alpha_2 x_3 + \beta_2,$$

где

$$\alpha_2 = -c_2/(b_2 + a_2 \alpha_1), \quad \beta_2 = (d_2 - a_2 \beta_1)/(b_2 + a_2 \alpha_1).$$

Продолжая процесс далее, на i -м ($1 < i < n$) шаге этого процесса i -е уравнение системы преобразуется к виду

$$x_i = \alpha_i x_{i+1} + \beta_i, \quad (2.19)$$

где

$$\alpha_i = -c_i/(b_i + a_i \alpha_{i-1}), \quad \beta_i = (d_i - a_i \beta_{i-1})/(b_i + a_i \alpha_{i-1}). \quad (2.20)$$

При $i = n$ из последнего уравнения системы находим

$$x_n = (d_n - a_n \beta_{n-1}) / (b_n + a_n \alpha_{n-1}) = \beta_n. \quad (2.21)$$

Полученные рекуррентные формулы позволяют решить систему (2.17) в два этапа. Прямой ход метода прогонки состоит в вычислении *прогоночных коэффициентов* α_i и β_i ($1 \leq i \leq n$) по формулам (2.18), (2.20) и (2.21).

Обратный ход прогонки состоит в вычислении неизвестных x_i . Сначала полагают $x_n = \beta_n$. Затем значения остальных неизвестных последовательно вычисляют по формуле (2.19).

Замечание 6. Если трехдиагональная матрица удовлетворяет условиям диагонального преобладания, то прогоночные коэффициенты $|\alpha_i| \leq 1$, и обратная прогонка устойчива по входным данным.

2.4. Метод Якоби

Итерационные методы применяются, главным образом, для решения задач большой размерности. Большие системы уравнений, возникающие в приложениях, как правило, являются разреженными.

Преобразуем систему $A \mathbf{x} = \mathbf{c}$ к виду, пригодному для итераций

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

или

$$x_i = \sum_{j=1}^n b_{ij} x_j + c_i, \quad 1 \leq i \leq n. \quad (2.22)$$

Правая часть определяет отображение F :

$$F: y_i = \sum_{j=1}^n b_{ij} x_j + c_i,$$

преобразующее точку $\mathbf{x} = (x_1, x_2, \dots, x_n)$ n -мерного векторного пространства в точку $\mathbf{y} = (y_1, y_2, \dots, y_n)$ того же пространства. Выбрав начальную точку $\mathbf{x}^0 = (x_1^0, x_2^0, \dots, x_n^0)$, можно построить итерационную последовательность $\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3, \dots, \mathbf{x}^n, \dots$ точек n -мерного векторного пространства аналогично методу простой итерации для скалярного уравнения $x = \varphi(x)$:

$$\mathbf{x}^{n+1} = F(\mathbf{x}^n). \quad (2.23)$$

При определенных условиях последовательность (2.23) сходится, и ее предел является решением системы (2.22). Напомним некоторые определения из курса математического анализа.

Определения:

1. Если в n -мерном векторном пространстве ввести одну из норм (2.3), позволяющую определить расстояние между произвольными точками \mathbf{x} и \mathbf{y} , то пространство будет *метрическим*.

$$\begin{aligned} \rho_1 = \|\mathbf{x} - \mathbf{y}\|_1 &= \sum_{i=1}^{i=n} |x_i - y_i|, & \rho_2 = \|\mathbf{x} - \mathbf{y}\|_2 &= \left(\sum_{i=1}^{i=n} (x_i - y_i)^2 \right)^{1/2}, \\ \rho_3 = \|\mathbf{x} - \mathbf{y}\|_\infty &= \max_{1 \leq i \leq n} |x_i - y_i|. \end{aligned} \quad (2.24)$$

2. Если в метрическом пространстве любая фундаментальная последовательность сходится, то пространство называется *полным*.

3. Пусть F – отображение, действующее в метрическом пространстве E с метрикой ρ , тогда отображение называется *сжимающим*, если существует такое число α , $0 < \alpha < 1$, что для любых двух точек $\mathbf{x}, \mathbf{y} \in E$ выполняется неравенство

$$\rho(F(x), F(y)) \leq \alpha \rho(x, y).$$

Аналогично теореме 1.1 для одномерного случая доказывается следующая теорема.

Теорема. Если F – сжимающее отображение, определенное в полном метрическом пространстве, то существует единственная неподвижная точка \mathbf{x} , такая, что $\mathbf{x} = F(\mathbf{x})$. При этом итерационная последовательность, построенная для отображения F с любым начальным приближением \mathbf{x}^0 , сходится к \mathbf{x} .

Имеет место критерий окончания итерационного процесса, аналогичный (1.4). Если значение α известно, то вычисление корня с точностью ε следует вести до выполнения неравенства

$$\frac{\alpha}{1 - \alpha} \rho(\mathbf{x}^n - \mathbf{x}^{n-1}) < \varepsilon,$$

или равносильного ему неравенства

$$\rho(\mathbf{x}^n, \mathbf{x}^{n-1}) < \frac{1 - \alpha}{\alpha} \varepsilon. \quad (2.25)$$

Заметим, что матрица коэффициентов системы имеет диагональное преобладание.

Приведем исходную систему к виду $\mathbf{x} = B\mathbf{x} + \mathbf{c}$, пригодному для итераций, для этого в каждом i -м уравнении исходной системы выразим x_i .

$$B = \begin{pmatrix} 0 & -0,16667 & -0,22222 & -0,11111 \\ -0,015789 & 0 & -0,47368 & 0,26316 \\ -0,17391 & -0,3913 & 0 & 0,30435 \\ -0,08333 & 0,20833 & 0,29167 & 0 \end{pmatrix}; \quad C = \begin{pmatrix} 1,11111 \\ -1,31579 \\ -4,86957 \\ 6,83333 \end{pmatrix}.$$

1. В пространстве с метрикой $\rho_1 = \|\mathbf{x} - \mathbf{y}\|_1 = \sum_{i=1}^{i=n} |x_i - y_i|$:

$$\varepsilon = \max_{1 \leq j \leq n} \sum_{i=1}^{i=n} |b_{ij}| = 0,89474 < 1.$$

2. В пространстве с метрикой $\rho_2 = \|\mathbf{x} - \mathbf{y}\|_2 = \left(\sum_{i=1}^{i=n} (x_i - y_i)^2 \right)^{1/2}$:

$$\alpha = \sqrt{\sum_{j,i=1}^n |b_{ij}|^2} = 0,90525 < 1.$$

3. В пространстве с метрикой $\rho_3 = \|\mathbf{x} - \mathbf{y}\|_\infty = \max_{1 \leq i \leq n} |x_i - y_i|$:

$$\alpha = \max_{1 \leq i \leq n} \sum_{j=1}^{j=n} |b_{ij}| = 0,98757 < 1.$$

Выберем метрику

$$\rho_2 = \|\mathbf{x} - \mathbf{y}\|_2 = \left(\sum_{i=1}^{i=n} (x_i - y_i)^2 \right)^{1/2} \text{ с } \alpha = 0,905251.$$

В качестве начального приближения $\mathbf{x}^0 = \text{colon}(20, -25, -112, 164)$ возьмем столбец из свободных членов исходной матрицы. Вычисление корня с точностью ε следует вести до выполнения неравенства

$$\rho_2(\mathbf{x}^n, \mathbf{x}^{n-1}) = \left(\sum_{i=1}^4 (x_i^n - x_i^{n-1})^2 \right)^{1/2} < \frac{1 - \alpha}{\alpha} \varepsilon$$

по формуле

$$x_i^n = \sum_{j=1}^4 b_{ij} x_j^{n-1} + c_i, \quad i = 1, 2, 3, 4; \quad n = 1, 2, \dots$$

На 53-м шаге итерационного процесса имеем

$$\mathbf{x}^{53} = \begin{pmatrix} 0,999865 \\ 2,000777 \\ -4,00311 \\ 5,99726 \end{pmatrix},$$

$$\rho_2(\mathbf{x}^{53}, \mathbf{x}^{52}) = \left(\sum_{i=1}^4 (x_i^{53} - x_i^{52})^2 \right)^{1/2} = 1,55E - 06 > \frac{1 - \alpha}{\alpha} \varepsilon.$$

На 54-м шаге итерационного процесса имеем

$$\mathbf{x}^{54} = \begin{pmatrix} 0,999865 \\ 2,000777 \\ -4,00311 \\ 5,99726 \end{pmatrix},$$

$$\rho_2(\mathbf{x}^{54}, \mathbf{x}^{53}) = \left(\sum_{i=1}^4 (x_i^{54} - x_i^{53})^2 \right)^{1/2} = 1,08E - 06 < \frac{1 - \alpha}{\alpha} \varepsilon.$$

Тогда решение системы с точностью $\varepsilon = 0,00001$:

$$\mathbf{x} \approx \mathbf{x}^{54} = \begin{pmatrix} 0,99987 \\ 2,00078 \\ -4,00311 \\ 5,99726 \end{pmatrix}.$$

2.5. Метод Зейделя

Основная идея метода Зейделя состоит в том, что на каждом шаге итерационного процесса при вычислении очередного $(k + 1)$ -го приближения учитываются уже найденные на этом этапе итерации приближения к неизвестным.

Пусть система $A \mathbf{x} = \mathbf{b}$ преобразована к виду (2.29), пригодному для итераций

$$\mathbf{x} = B\mathbf{x} + \mathbf{c}, \quad b_{ij} = -\frac{a_{ij}}{a_{ii}}, \quad c_i = \frac{b_i}{a_{ii}}.$$

На $(k+1)$ -й итерации компоненты приближения \mathbf{x}^{k+1} вычисляются по формулам

$$x_1^{k+1} = \sum_{j=1}^n b_{1j} x_j^k + c_1, \quad x_i^{k+1} = \sum_{j=1}^{i-1} b_{ij} x_j^{k+1} + \sum_{j=i}^n b_{ij} x_j^k + c_i, \quad 2 \leq i \leq n. \quad (2.30)$$

Введем нижнюю и верхнюю треугольные матрицы

$$T_H = \begin{bmatrix} 0 & 0 & 0 & 0 \\ b_{21} & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots \\ b_{n1} & \dots & b_{nn-1} & 0 \end{bmatrix}, \quad T_B = \begin{bmatrix} 0 & b_{12} & \dots & b_{1n} \\ 0 & 0 & \dots & b_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

тогда матрица $B = T_H + T_B$.

Расчетные формулы метода в компактном виде:

$$\mathbf{x}^{k+1} = T_H \mathbf{x}^{k+1} + T_B \mathbf{x}^k + \mathbf{c}. \quad (2.31)$$

Достаточные условия сходимости (2.26)–(2.28) являются достаточными условиями сходимости метода Зейделя. В случае, когда матрица A системы A – симметричная и положительная, имеет место следующая теорема.

Теорема. Пусть A – симметричная и положительная матрица. Тогда при любом выборе начального приближения $\mathbf{x}^0 = (x_1^0, x_2^0, \dots, x_m^0)$ метод Зейделя сходится со скоростью геометрической прогрессии.

Замечание 7. В случае, когда матрица A системы A обладает свойством диагонального преобладания или является симметричной с положительными диагональными элементами, имеет место аналогичная теорема.

Систему (2.1) всегда можно привести к виду, для которой метод Зейделя сходится. Умножая обе части уравнения (2.1) на A^T , получим

$$A^T A \mathbf{x} = A^T \mathbf{b},$$

где $C = A^T A$ есть диагональная матрица с положительными диагональными элементами.

Геометрическая интерпретация метода в случае решения системы двух уравнений методом Зейделя приведена на рис. 12.

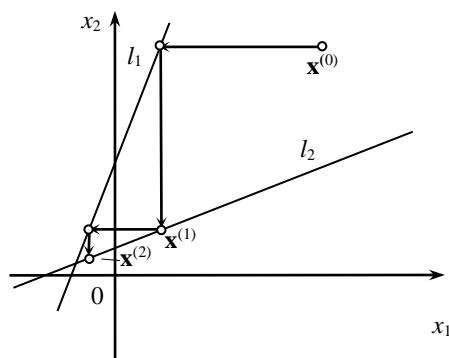


Рис. 12

Пример 2. Решить систему уравнений методом Зейделя

$$\begin{aligned} 18x_1 + 3x_2 + 4x_3 + 2x_4 &= 20, \\ 3x_1 + 19x_2 + 9x_3 - 5x_4 &= -25, \\ 4x_1 + 9x_2 + 23x_3 - 7x_4 &= -112, \\ 2x_1 - 5x_2 - 7x_3 + 24x_4 &= 164. \end{aligned}$$

с точностью $\varepsilon = 0,00001$.

Заметим, что матрица коэффициентов системы имеет диагональное преобладание.

Приведем исходную систему к виду $\mathbf{x} = B\mathbf{x} + \mathbf{c}$, пригодному для итераций, для этого в каждом i -м уравнении исходной системы выразим x_i .

$$B = \begin{pmatrix} 0 & -0,16667 & -0,22222 & -0,11111 \\ -0,015789 & 0 & -0,47368 & 0,26316 \\ -0,17391 & -0,3913 & 0 & 0,30435 \\ -0,08333 & 0,20833 & 0,29167 & 0 \end{pmatrix}; \quad C = \begin{pmatrix} 1,11111 \\ -1,31579 \\ -4,86957 \\ 6,83333 \end{pmatrix}.$$

Как и в методе Якоби выберем метрику

$$\rho_2 = \|\mathbf{x} - \mathbf{y}\|_2 = \left(\sum_{i=1}^4 (x_i - y_i)^2 \right)^{1/2} \quad \text{с } \alpha = 0,905251.$$

В качестве начального приближения \mathbf{x}^0 возьмем столбец *colon*(20, -25, -112, 164) из свободных членов исходной системы линейных уравнений.

Вычисление корня с точностью ε следует вести до выполнения критерия

$$\rho_2(\mathbf{x}^n, \mathbf{x}^{n-1}) = \left(\sum_{i=1}^4 (x_i^n - x_i^{n-1})^2 \right)^{1/2} < \frac{1-\alpha}{\alpha} \varepsilon,$$

по формулам

$$x_1^{k+1} = \sum_{j=1}^4 b_{1j} x_j^k + c_1,$$

$$x_i^{k+1} = \sum_{j=1}^{i-1} b_{ij} x_j^{k+1} + \sum_{j=i}^4 b_{ij} x_j^k + c_i, \quad i = 2, 3, 4; \quad k = 0, 1, 2, \dots$$

На 13-м шаге итерационного процесса имеем

$$\mathbf{x}^{13} = \begin{pmatrix} 0,999864 \\ 2,000779 \\ -4,003115 \\ 5,997265 \end{pmatrix},$$

$$\rho_2(\mathbf{x}^{13}, \mathbf{x}^{12}) = \left(\sum_{i=1}^4 (x_i^{13} - x_i^{12})^2 \right)^{1/2} = 2,27E - 06 > \frac{1-\alpha}{\alpha} \varepsilon.$$

На 14-м шаге итерационного процесса получили

$$\mathbf{x}^{14} = \begin{pmatrix} 0,999864 \\ 2,000777 \\ -4,0031117 \\ 5,9972642 \end{pmatrix},$$

$$\rho_2(\mathbf{x}^{14}, \mathbf{x}^{13}) = \left(\sum_{i=1}^4 (x_i^{14} - x_i^{13})^2 \right)^{1/2} = 4,39E - 07 < \frac{1-\alpha}{\alpha} \varepsilon.$$

Решение системы с точностью $\varepsilon = 0,00001$ достигается уже на 14-м шаге итераций

$$\mathbf{x} \approx \mathbf{x}^{14} = \begin{pmatrix} 0,99986 \\ 2,00078 \\ -4,00311 \\ 5,99726 \end{pmatrix}.$$

2.6. Двухслойные итерационные методы

Двухслойные итерационные методы могут быть записаны в *канонической форме*

$$B_{k+1} \frac{\mathbf{x}^{k+1} - \mathbf{x}^k}{\tau_{k+1}} + A\mathbf{x}^k = \mathbf{b}, \quad k = 0, 1, 2, \dots \quad (2.32)$$

Здесь B_{k+1} – некоторая невырожденная матрица, $\tau_{k+1} > 0$ – итерационные параметры. В случае, когда параметры B_{k+1} и τ_{k+1} не зависят от номера итерации, метод называют *стационарным*.

Если $B = E$ – единичной матрице, то итерационный метод называется *явным*, поскольку в нем очередное приближение явным образом выражается через предыдущее:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \tau_{k+1} (A\mathbf{x}^k - \mathbf{b}),$$

или

$$x_i^{k+1} = x_i^k - \tau_{k+1} \left(\sum_{j=1}^n a_{ij} x_j^k - b_i \right).$$

В общем случае, при $B \neq E$, метод является *неявным*, т. к. на каждой итерации требуется решать систему уравнений

$$B\mathbf{x}^{k+1} = B\mathbf{x}^k - \tau_{k+1} (A\mathbf{x}^k - \mathbf{b}).$$

Выбирают простые матрицы B . Например, диагональные, трехдиагональные, треугольные.

Метод Якоби соответствует выбору $B = D$ – диагональной матрице и $\tau_{k+1} = 1$.

Метод Зейделя (2.30) является неявным методом и соответствует выбору $B = D + T_H$, и $\tau_{k+1} = 1$.

Для стационарного итерационного процесса приведем без доказательства достаточное условие сходимости.

Теорема Самарского. Пусть A – симметричная и положительно определенная матрица; $B - \frac{\tau}{2}A$ – положительно определенная матрица; τ – положительное число. Тогда при любом выборе начального приближения \mathbf{x}^0 итерационный процесс, который определяется рекуррентной формулой (2.30), сходится к решению исходной системы (2.1).

Определим интервал, в котором может меняться параметр τ :

$$(B\mathbf{x}, \mathbf{x}) > \frac{\tau}{2}(A\mathbf{x}, \mathbf{x}).$$

Откуда

$$0 < \tau < \tau_0 = \min_{\mathbf{x} \neq 0} \frac{2(B\mathbf{x}, \mathbf{x})}{(A\mathbf{x}, \mathbf{x})}. \quad (2.33)$$

Рассмотрим *метод простой итерации* ($B = E$) с параметром τ в канонической форме:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \tau(A\mathbf{x}^k - \mathbf{b}). \quad (2.34)$$

Будем считать, что A – симметричная и положительно определенная матрица, тогда из (2.33) и замечания 4, находим верхнюю границу интервала сходимости по итерационному параметру:

$$\tau_0 = \min_{\mathbf{x} \neq 0} \frac{2(\mathbf{x}, \mathbf{x})}{(A\mathbf{x}, \mathbf{x})} = \min_{\mathbf{x} \neq 0} \frac{2}{\frac{(A\mathbf{x}, \mathbf{x})}{(\mathbf{x}, \mathbf{x})}} = \frac{2}{\max_{\mathbf{x} \neq 0} \frac{(A\mathbf{x}, \mathbf{x})}{(\mathbf{x}, \mathbf{x})}} = \frac{2}{\lambda_{\max}}.$$

Пример 3. Решить систему уравнений из прим. 2 методом простой итерации с выбором параметра τ .

При помощи одной из стандартных программ найдем собственные значения матрицы A : $\lambda_1 = 11,69064$, $\lambda_2 = 13,37067$, $\lambda_3 = 22,20059$, $\lambda_4 = 36,73809$.

Параметр τ в силу (2.33) удовлетворяет неравенству $0 < \tau < \tau_0 = \frac{2}{\lambda_{\max}} = 0,05444$. Возьмем $\tau = \tau_0/2 = 0,02722$ и будем искать каждое очередное приближение решения исходной системы уравнений по формуле

$$\mathbf{x}^{k+1} = \mathbf{x}^k - 0,02722(A\mathbf{x}^k - \mathbf{b}),$$

где

$$A = \begin{pmatrix} 18 & 3 & 4 & 2 \\ 3 & 19 & 9 & -5 \\ 4 & 9 & 23 & -7 \\ 2 & -5 & -7 & 24 \end{pmatrix}; \quad b = \begin{pmatrix} 20 \\ -25 \\ -112 \\ 164 \end{pmatrix}.$$

Выберем метрику $\rho_2 = \|\mathbf{x} - \mathbf{y}\|_2 = \left(\sum_{i=1}^{i=n} (x_i - y_i)^2\right)^{1/2}$ с $\alpha = 0,905251$.

В качестве начального приближения \mathbf{x}^0 возьмем столбец \mathbf{b} из свободных членов исходной системы уравнений. Вычисление корня с точностью ε следует вести до выполнения неравенства

$$\rho_2(\mathbf{x}^n, \mathbf{x}^{n-1}) = \left(\sum_{i=1}^4 (x_i - y_i)^2\right)^{1/2} < \frac{1 - \alpha}{\alpha} \varepsilon.$$

На 38-м шаге итерационного процесса имеем

$$\mathbf{x}^{38} = \begin{pmatrix} 1,000002 \\ 2,000013 \\ -4,000010 \\ 5,999998 \end{pmatrix},$$

$$\rho_2(\mathbf{x}^{38}, \mathbf{x}^{37}) = \left(\sum_{i=1}^4 |x_i^{38} - x_i^{37}|^2\right)^{1/2} = 1,335749 E - 05 > \frac{1 - \alpha}{\alpha} \varepsilon.$$

На 39-м шаге итерационного процесса имеем

$$\mathbf{x}^{39} = \begin{pmatrix} 1,000001 \\ 2,000009 \\ -4,000001 \\ 5,999999 \end{pmatrix},$$

$$\rho_2(\mathbf{x}^{39}, \mathbf{x}^{38}) = \left(\sum_{i=1}^4 |x_i^{39} - x_i^{38}|^2\right)^{1/2} = 9,03022 E - 06 < \frac{1 - \alpha}{\alpha} \varepsilon.$$

Решение системы с точностью $\varepsilon = 0,00001$ достигается на 39-м шаге итераций

$$\mathbf{x} \approx \mathbf{x}^{39} = \begin{pmatrix} 1,00000 \\ 2,00000 \\ -4,00001 \\ 5,99999 \end{pmatrix}.$$

Скорость сходимости метода простой итерации оказалась существенно ниже скорости сходимости метода Зейделя, но выше метода Якоби при решении одной и той же системы линейных уравнений при одинаковом начальном приближении.

Модифицируем метод Зейделя. С этой целью введем параметр ω и запишем рекуррентное соотношение (2.32) метода верхней релаксации в виде

$$(D + \omega T_H) \frac{\mathbf{x}^{k+1} - \mathbf{x}^k}{\omega} + A\mathbf{x}^k = \mathbf{b},$$

или

$$\left(\frac{D}{\omega} + \hat{D}_f \right) (\mathbf{x}^{k+1} - \mathbf{x}^k) + A\mathbf{x}^k = \mathbf{b}. \quad (2.35)$$

В данном случае

$$B = D + \omega T_H, \quad \tau_k = \omega > 0.$$

Подставляя вместо матрицы A выражение $A = T_H + D + T_B$ в (2.35), получим

$$\left(\frac{D}{\omega} + \hat{D}_f \right) \mathbf{x}^{k+1} + \mathbf{x}^k \left((1 - 1/\omega)D + T_{\hat{A}} \right) = \mathbf{b}.$$

Перейдем от векторной записи к записи в виде отдельных уравнений:

$$x_i^{k+1} = x_i^k + \frac{\omega}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{k+1} - \sum_{j=i+1}^n a_{ij} x_j^k \right), \quad i = 1, 2, \dots, n. \quad (2.36)$$

При $\omega = 1$ мы возвращаемся к методу Зейделя (2.30).

Выведем достаточное условие для сходимости метода верхней релаксации (2.35) из теоремы Самарского:

$$B - \frac{\tau}{2} A = (D + \omega \hat{D}_f) - \frac{\omega}{2} (D + \hat{D}_f + \hat{D}_{\hat{A}}) = \left(1 - \frac{\omega}{2} \right) D + \frac{\omega}{2} (\hat{D}_f - \hat{D}_{\hat{A}}).$$

Из симметричности и положительности матрицы A соответственно следует:

$$(T_H \mathbf{x}, \mathbf{x}) = (T_H^T \mathbf{x}, \mathbf{x}) = (T_B \mathbf{x}, \mathbf{x}), \quad (D\mathbf{x}, \mathbf{x}) > 0.$$

Поэтому неравенство

$$\left(\left(B - \frac{\tau}{1} A \right) \mathbf{x}, \mathbf{x} \right) = \left(\left(1 - \frac{\omega}{1} \right) D \mathbf{x}, \mathbf{x} \right) > 0$$

выполняется при $0 < \omega < 2$.

Условие теоремы Самарского выполняется, когда параметр метода $0 < \omega < 2$.

Метод Зейделя, соответствующий случаю $\omega = 1$, сходится, тем самым доказана теорема 2.2 и замечание к ней.

Если сравнить метод Зейделя с методом верхней релаксации, то суть метода состоит в следующем. На $(k + 1)$ -й итерации компоненты приближения \mathbf{x}^{k+1} вычисляются по формулам метода Зейделя:

$$\tilde{x}_i^{k+1} = x_j^k + \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{k+1} - \sum_{j=i}^n a_{ij} x_j^k \right),$$

а затем производят дополнительно смещение на величину $(\omega - 1)(\tilde{x}_i^{k+1} - x_i^k)$.

Таким образом, компонента $(k + 1)$ -го приближения вычисляется по формуле

$$x_i^{k+1} = \tilde{x}_i^{k+1} + (\omega - 1)(\tilde{x}_i^{k+1} - x_i^k) = \omega \tilde{x}_i^{k+1} + (1 - \omega)x_i^k.$$

Геометрическая интерпретация метода верхней релаксации в случае решения системы двух уравнений приведена на рис. 13.

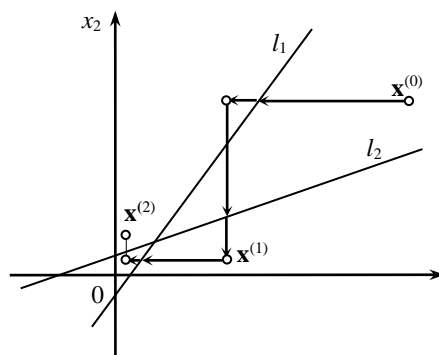


Рис. 13

Пример 4. Решить систему уравнений из примера 2 методом верхней релаксации с точностью $\varepsilon = 0,00001$.

Выберем метрику $\rho_2 = \|\mathbf{x} - \mathbf{y}\|_2 = \left(\sum_{i=1}^4 |x_i - y_i|^2 \right)^{1/2}$ с $\alpha = 0,905251$.

В качестве начального приближения \mathbf{x}^0 в итерационном методе верхней релаксации возьмем $colon(20, -25, -112, 164)$ из свободных членов исходной системы уравнений.

Вычисление корня с точностью ε следует вести до выполнения неравенства

$$\rho_2(\mathbf{x}^n, \mathbf{x}^{n-1}) = \left(\sum_{i=1}^4 |x_i^n - x_i^{n-1}|^2 \right)^{1/2} < \frac{1-\alpha}{\alpha} \varepsilon,$$

по формулам

$$x_i^{k+1} = x_i^k + \frac{\omega}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{k+1} - \sum_{j=i}^4 a_{ij} x_j^k \right), \quad i=1, 2, 3, 4; \quad k=0, 1, 2, \dots$$

Выберем параметр $\omega = 0,5$.

На 37-м шаге итерационного процесса имеем

$$\mathbf{x}^{37} = \begin{pmatrix} 0,9999964 \\ 1,999995 \\ -3,999999 \\ 6,000005 \end{pmatrix},$$

$$\rho_2(\mathbf{x}^{37}, \mathbf{x}^{36}) = \left(\sum_{i=1}^4 |x_i^{37} - x_i^{36}|^2 \right)^{1/2} = 1,07E-05 > \frac{1-\alpha}{\alpha} \varepsilon.$$

На 38-м шаге итерационного процесса получили

$$\mathbf{x}^{38} = \begin{pmatrix} 0,9999966 \\ 1,999996 \\ -3,999999 \\ 6,000003 \end{pmatrix},$$

$$\rho_2(\mathbf{x}^{38}, \mathbf{x}^{37}) = \left(\sum_{i=1}^4 |x_i^{38} - x_i^{37}|^2 \right)^{1/2} = 7,42E-06 < \frac{1-\alpha}{\alpha} \varepsilon.$$

Решение системы с точностью $\varepsilon = 0,00001$ достигается на 38-м шаге итерационного приближения

$$\mathbf{x} \approx \mathbf{x}^{38} = \begin{pmatrix} 1,00000 \\ 2,00000 \\ -3,99999 \\ 6,00000 \end{pmatrix}.$$

Скорость сходимости метода верхней релаксации при выборе параметра $\omega = 0,5$ оказалась существенно ниже скорости сходимости метода Зейделя, т. е. не произошло улучшение метода Зейделя за счет введения параметра $\omega = 0,5$.

В предположении, что матрица Якоби невырожденная, получаем итерационную формулу метода Ньютона:

$$\mathbf{f}(\mathbf{x}^n) + \mathbf{f}'(\mathbf{x}^n)(\mathbf{x}^{n+1} - \mathbf{x}^n) = 0; \quad \mathbf{x}^{n+1} = \mathbf{x}^n - (\mathbf{f}'(\mathbf{x}^n))^{-1} \mathbf{f}(\mathbf{x}^n). \quad (3.9)$$

Формула (3.9) предполагает обращение матрицы Якоби, что является довольно трудоемкой операцией. Поэтому сначала решают систему линейных алгебраических уравнений

$$\mathbf{f}'(\mathbf{x}^n) \Delta \mathbf{x}^{n+1} = -\mathbf{f}(\mathbf{x}^n) \quad (3.10)$$

относительно поправки $\Delta \mathbf{x}^{n+1} = (\mathbf{x}^{n+1} - \mathbf{x}^n)$.

Затем определяют

$$\mathbf{x}^{n+1} = \Delta \mathbf{x}^{n+1} + \mathbf{x}^n. \quad (3.11)$$

Метод Ньютона сходится с квадратичной скоростью и имеет простой критерий окончания итераций

$$\|\mathbf{x}^n - \mathbf{x}^{n-1}\| \leq \varepsilon, \quad (3.12)$$

как и в случае одного нелинейного уравнения.

Замечание 1. Использование метода Ньютона имеет сложность, которая, во-первых, заключается в вычислении на каждом шаге итерации матрицы Якоби, содержащей m^2 частных производных. Во вторых, существует трудность в нахождении начального приближения.

Пример 1. Решить методом Ньютона систему уравнений

$$\begin{aligned} x^2 + 2xy - 3 &= 0, \\ 2x^2 - 3xy + 1 &= 0. \end{aligned}$$

Составим матрицу Якоби

$$\mathbf{f}'(x, y) = \begin{pmatrix} 2x + 2y & 2x \\ 4x - 3y & -3x \end{pmatrix}.$$

В качестве начального приближения возьмем точку $\mathbf{x}^0 = (0,8; 0,8)$.

Тогда уравнение (3.10) примет вид

$$\begin{aligned} \mathbf{f}'(0,8;0,8)\Delta\mathbf{x}^1 &= -\mathbf{f}(0,8;0,8), \\ 3,2\Delta x^1 + 1,6\Delta y^1 &= 1,08, \\ 0,8\Delta x^1 - 2,4\Delta y^1 &= -0,36. \end{aligned}$$

Откуда $\Delta x^1 = 0,225$, $\Delta y^1 = 0,225$.

Тогда $x^1 = 0,8 + 0,225 = 1,025$, $y^1 = 1,025$.

Следующую поправку $\Delta\mathbf{x}^2 = (\Delta x^2, \Delta y^2)$ находим из уравнения

$$\mathbf{f}'(1,025;1,025)\Delta\mathbf{x}^2 = -\mathbf{f}(1,025;1,025),$$

или

$$\begin{aligned} 4,1\Delta x^2 + 1,05\Delta y^2 &= -0,151875, \\ 1,025\Delta x^2 - 3,075\Delta y^2 &= 0,050625. \end{aligned}$$

Откуда $\Delta x^2 = -0,029813$, $\Delta y^2 = -0,028230$.

Тогда

$$\begin{aligned} x^2 &= 1,0250 - 0,029813 = 0,995819; \\ y^2 &= 1,025 - 0,028230 = 0,996770. \end{aligned}$$

Следующую поправку $\Delta\mathbf{x}^3 = (\Delta x^3, \Delta y^3)$ находим из уравнения

$$\mathbf{f}'(0,995819;0,996770)\Delta\mathbf{x}^3 = -\mathbf{f}(0,995819;0,996770),$$

Откуда $\Delta x^3 = 0,0005803$, $\Delta y^3 = 0,000067$.

Тогда

$$\begin{aligned} x^3 &= 0,995819 + 0,0005803 = 0,996349, \\ y^3 &= 0,996770 + 0,000067 = 0,996837. \end{aligned}$$

Легко проверить, что решение данной системы $x = y = 1$. Из приведенных итераций видно, что метод Ньютона быстро сходится при хорошем начальном приближении.

3.3. Модификации метода Ньютона для решения системы нелинейных уравнений

Если в расчетной формуле (3.10) заменить матрицу $\mathbf{f}'(\mathbf{x}^n)$ постоянной матрицей $A = \mathbf{f}'(\mathbf{x}^0)$, то получим формулы *упрощенного метода Ньютона*

$$A\Delta\mathbf{x}^{n+1} = -\mathbf{f}(\mathbf{x}^n), \quad \mathbf{x}^{n+1} = \Delta\mathbf{x}^{n+1} + \mathbf{x}^n. \quad (3.13)$$

По сравнению с методом Ньютона число итераций, необходимое для достижения заданной точности, существенно возрастает.

Элементы матрицы $\mathbf{f}'(\mathbf{x}^n)$ приближенно можно вычислять с помощью формул конечно-разностной аппроксимации частной производной:

$$\begin{aligned} \frac{\partial f_i(\mathbf{x}^n)}{\partial x_j} \approx J_{ij}^n = \frac{1}{h_j^n} (f_i(x_1^n, \dots, x_{j-1}^n, x_j^n + h_j^n, x_{j+1}^n, \dots, x_m^n) - \\ - f_i(x_1^n, \dots, x_{j-1}^n, x_j^n, x_{j+1}^n, \dots, x_m^n)). \end{aligned} \quad (3.14)$$

Если в формуле (3.10) матрицу $\mathbf{f}'(\mathbf{x}^n)$ заменить аппроксимирующей ее матрицей (J_{ij}^n) по формуле (3.14) с шагом $\mathbf{h}^n = \mathbf{x}^{n-1} - \mathbf{x}^n$, то получим *метод секущих*:

$$J\Delta\mathbf{x}^{n+1} = -\mathbf{f}(\mathbf{x}^n), \quad \mathbf{x}^{n+1} = \Delta\mathbf{x}^{n+1} + \mathbf{x}^n. \quad (3.15)$$

Для того чтобы начать вычисления, необходимо задать два начальных приближения $\mathbf{x}^0, \mathbf{x}^1$.

Если матрицу $\mathbf{f}'(\mathbf{x}^n)$ заменить аппроксимирующей ее матрицей (J_{ij}^n) по формуле (5.14) с шагом $\mathbf{h}^n = \mathbf{f}(\mathbf{x}^n)$, то получим *метод Стеффенсена*. Этот метод обладает свойством квадратичной сходимости и является одношаговым, как и метод Ньютона.

3.4. Методы спуска для решения системы нелинейных уравнений

Иногда эффективным способом решения системы нелинейных уравнений является сведение к задаче отыскания минимума функции многих переменных. Введем функцию

$$\hat{O}(\mathbf{x}) = f_1^2(x_1, x_2, \dots, x_m) + f_2^2(x_1, x_2, \dots, x_m) + \dots + f_m^2(x_1, x_2, \dots, x_m). \quad (3.16)$$

Она неотрицательна и достигает своего минимума тогда и только тогда, когда

$$\begin{aligned} f_1(x_1, x_2, \dots, x_m) &= 0, \\ f_2(x_1, x_2, \dots, x_m) &= 0, \\ &\dots\dots\dots \\ f_m(x_1, x_2, \dots, x_m) &= 0. \end{aligned}$$

Рассмотрим один из итерационных методов минимизации функции многих переменных – *метод спуска*. Методы спуска, как правило, имеют более широкую область сходимости, чем рассмотренные ранее методы. Для простой графической иллюстрации метода спуска рассмотрим случай двух нелинейных уравнений, тогда целевая функция $\Phi(x_1, x_2) = f_1^2(x_1, x_2) + f_2^2(x_1, x_2)$. Множество точек, для которых $\Phi(x_1, x_2) = C$, называется *поверхностью уровня*.

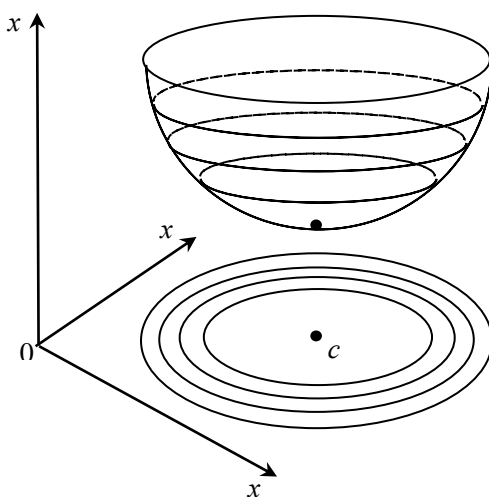


Рис. 14

В трехмерном пространстве функция $x_3 = \Phi(x_1, x_2)$ задает некоторую поверхность, низшая точка которой и дает решение задачи минимизации. Если провести плоскости $x_3 = C$, то проекции на плоскость Ox_1x_2 линий пересечения этих плоскостей с поверхностями дают *линии уровня* (поверхности уровня) (рис. 14).

Как известно, $\mathbf{grad} \hat{O}(x_1, x_2) = \frac{\partial \hat{O}}{\partial x_1} \mathbf{i} + \frac{\partial \hat{O}}{\partial x_2} \mathbf{j}$ перпендикулярен поверхности уровня в точке $\mathbf{x} = (x_1, x_2)$ и указывает направление наискорейшего возрастания функции $\Phi(x_1, x_2)$. Вектор $\mathbf{p} = -\mathbf{grad} \hat{O}(x_1, x_2)$ называется *антиградиентом* и указывает направление наискорейшего убывания функции (рис. 15).

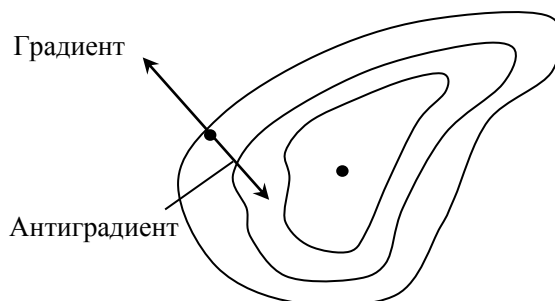


Рис. 15

При построении итерационной последовательности в задаче безусловной минимизации некоторой функции $\Phi(\mathbf{x})$ от n переменных строится последовательность $\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3, \dots, \mathbf{x}^n$, такая, что $\Phi(\mathbf{x}^{n+1}) < \Phi(\mathbf{x}^n)$. Кратко опишем шаги метода спуска для построения итерационной последовательности.

1. На каждом шаге итерации находят вектор \mathbf{p}^n — *направление спуска*, что при всех достаточно малых $\alpha > 0$ выполняется неравенство $\Phi(\mathbf{x}^n + \alpha \mathbf{p}^n) < \Phi(\mathbf{x}^n)$. Если ввести функцию одной переменной $\varphi_n(\alpha) = \Phi(\mathbf{x}^n + \alpha \mathbf{p}^n)$, то $\varphi_n(\alpha) < \varphi_n(0)$.

2. Вычисляют положительное число α_n — *шаг спуска*, для которого выполняется неравенство или $\varphi_n(\alpha_n) < \varphi_n(0)$.

3. За очередное приближение к точке минимума принимают $\mathbf{x}^{n+1} = \mathbf{x}^n + \alpha_n \mathbf{p}^n$.

4. Проверяют выполнение критерия окончания итераций. Если критерий выполняется, то итерации прекращаются и полагают $\mathbf{c} = \mathbf{x}^{n+1}$.

Замечание 2. Из неравенства $\varphi_n(\alpha_n) < \varphi_n(0)$ следует, что $\varphi'_n(0) < 0$. Так как $\varphi'_n(\alpha) = \mathbf{grad} \hat{O}(\mathbf{x}^n + \alpha \mathbf{p}^n) \mathbf{p}^n$, то вектор \mathbf{p}^n должен удовлетворять условию $\mathbf{grad} \hat{O}(\mathbf{x}^n) \mathbf{p}^n < 0$.

Выбор в качестве вектора \mathbf{p}^n антиградиента

$$\mathbf{p}^n = -\mathbf{grad} \hat{O}(\mathbf{x}^n) \quad (3.17)$$

задает *градиентный метод*.

В этом случае

$$\mathbf{grad} \hat{O}(\mathbf{x}^n) \mathbf{p}^n = -(\mathbf{grad} \hat{O}(\mathbf{x}^n))^2 < 0.$$

Шаг спуска α_n связан с «истинной» величиной шага h_n формулой $h_n = \|\mathbf{x}^{n+1} - \mathbf{x}^n\| = \alpha_n \|\mathbf{p}^n\|$. Одним из простейших методов определения шага α_n заключается в *дроблении шага* по формуле $\alpha_n = \alpha(1/2)^n$, где α – некоторое фиксированное положительное число, а n определяется из условия.

На практике используются следующие критерии окончания итераций:

$$\|\mathbf{x}^{n+1} - \mathbf{x}^n\| < \varepsilon_1, \quad (3.18)$$

$$\|\Phi(\mathbf{x}^{n+1}) - \Phi(\mathbf{x}^n)\| < \varepsilon_2, \quad (3.19)$$

$$\|\Phi'(\mathbf{x}^n)\| < \varepsilon_3. \quad (3.20)$$

Замечание 3. Известно, что градиентный метод сходится очень медленно, если поверхности уровня минимизируемой функции сильно вытянуты в некоторых направлениях. В двумерном случае линии уровня напоминают рельеф местности с оврагом.

В таких случаях для ускорения градиентного метода разработаны специальные методы. Сначала совершают градиентный спуск «на дно оврага» из двух соседних точек \mathbf{z}_0 и \mathbf{z}_1 (рис. 16). Через две найденные точки \mathbf{x}_0 и \mathbf{x}_1 проводят прямую, вдоль которой совершают большой шаг через «овраг». Из найденной точки \mathbf{z}_2 делают один градиентный спуск в точку \mathbf{x}_2 и совершают большой шаг вдоль прямой, проходящей через точки \mathbf{x}_1 и \mathbf{x}_2 , и т. д.

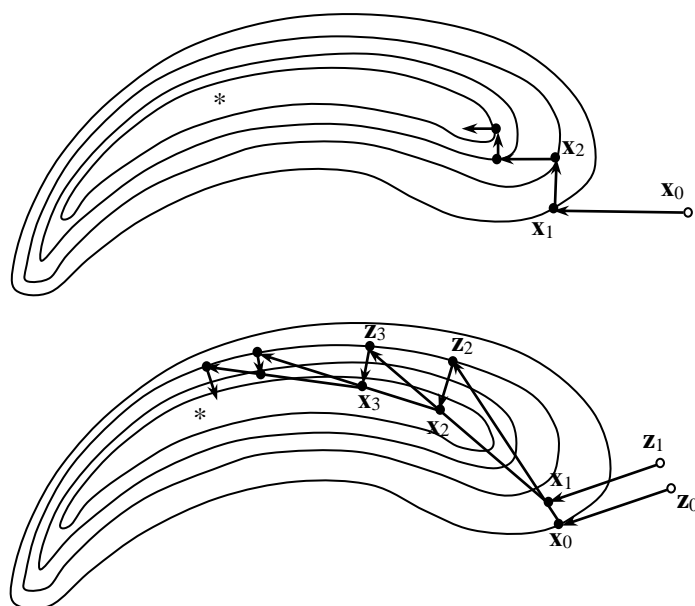


Рис. 16

Пример 2. Решить методом градиентного спуска систему уравнений

$$\begin{aligned}x^2 + 2xy - 3 &= 0, \\ 2x^2 - 3xy + 1 &= 0.\end{aligned}$$

Составляем целевую функцию:

$$\Phi(x, y) = (x^2 + 2xy - 3)^2 + (2x^2 - 3xy + 1)^2.$$

Вычислим градиент этой функции:

$$\begin{aligned}\mathbf{grad} \hat{O}(x, y) &= \frac{\partial \hat{O}}{\partial x} \mathbf{i} + \frac{\partial \hat{O}}{\partial y} \mathbf{j} = \\ &= 2((2x + 2y)(x^2 + 2xy - 3) + (4x - 3y)(2x^2 - 3xy + 1)) \mathbf{i} + \\ &\quad + 2(2x(x^2 + 2xy - 3) - 3x(2x^2 - 3xy + 1)) \mathbf{j}.\end{aligned}$$

В качестве начального приближения возьмем точку $\mathbf{x}^0 = (x^0, y^0) = (0,5; 0,5)$.

Тогда

$$\mathbf{p}^0 = -\mathbf{grad} \hat{O}(\mathbf{x}^0) = 8,25 \mathbf{i} + 6,75 \mathbf{j}.$$

Выберем шаг спуска α_0 из условия

$$\Phi(0,5 + 8,25\alpha_0; 0,5 + 6,75\alpha_0) < \Phi(0,5; 0,5).$$

Примем $\alpha_0 = 0,01$, тогда $\Phi(0,5825; 0,5675) = 4,470065$ и $\Phi(0,5; 0,5) = 5,625$. Поэтому можно положить $\mathbf{x}^1 = (0,5825; 0,5675)$.

Следующее направление спуска найдем по формуле

$$\mathbf{p}^1 = -\mathbf{grad} \hat{O}(\mathbf{x}^1) = 8,33589 \mathbf{i} + 7,05970 \mathbf{j}.$$

Возьмем $\alpha_1 = 0,01$, тогда $\Phi(0,66586; 0,63810) = 3,28805$. Поэтому можно положить $\mathbf{x}^2 = (0,66586; 0,63810)$.

Следующее направление спуска найдем по формуле

$$\mathbf{p}^2 = -\mathbf{grad} \hat{O}(\mathbf{x}^2) = 7,98563 \mathbf{i} + 6,99152 \mathbf{j}$$

и т. д.

При вычислениях методом спуска уже при небольшом количестве шагов происходит довольно быстрое приближение к решению системы, но затем скорость сходимости приближений к решению системы заметно уменьшается. За сорок итераций метода спуска с параметром $\alpha = 0,01$ получили менее точное решение $x = 1,00002$, $y = 0,99997$, чем при вычислении методом Ньютона за четыре итерации при хорошем начальном приближении.

На практике часто используют комбинацию методов: сначала метод спуска для определения начального приближения, а затем метод Ньютона для быстрого нахождения решения системы.

4. ПРИБЛИЖЕНИЕ ФУНКЦИЙ

Пусть на $[a,b]$ определена некоторая функция $y = f(x)$, однако полная информация о ней неизвестна. Известны лишь ее значения в конечном числе точек x_0, x_1, \dots, x_n , этого отрезка, которые будем считать занумерованными в порядке возрастания:

$$a \leq x_0 < x_1 < \dots < x_i < x_{i+1} < \dots < x_n \leq b.$$

Требуется по известным табличным значениям

$$y_i = f(x_i), \quad i = 0, 1, \dots, n; \tag{4.1}$$

«восстановить» исходную функцию $y = f(x)$, т. е. построить на отрезке $[a,b]$ функцию $F(x)$, достаточно близкую к $f(x)$. Функцию $F(x)$ называют *интерполирующей*, точки $x = x_0, x = x_1, \dots, x = x_n$ — *узлами интерполяции*.

Иногда, непосредственное вычисление функции $y = f(x)$ связано с проведением сложных расчетов. Нередко значения (4.1) находятся из эксперимента, и табличные значения y_i^* отличаются от «истинных» y_i , т. к. заведомо содержат ошибки эксперимента. В такой ситуации строят приближенную функцию $F(x)$, так, чтобы погрешность приближения была достаточно мала.

4.1. Интерполирование полиномами. Полином Лагранжа

Для заданных значений (4.1) многочлен $P_n(x) = \sum_{k=0}^n a_k x^k$ называется интерполяционным, если он удовлетворяет условиям

$$P_n(x_i) = y_i, \quad i = 0, 1, \dots, n; \tag{4.2}$$

Равенство (4.2) можно записать в виде системы уравнений

$$\begin{aligned} a_0 + a_1 x_0 + a_2 x_0^2 + \dots + a_n x_0^n &= y_0, \\ a_0 + a_1 x_1 + a_2 x_1^2 + \dots + a_n x_1^n &= y_1, \\ \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \\ a_0 + a_1 x_n + a_2 x_n^2 + \dots + a_n x_n^n &= y_n \end{aligned}$$

относительно коэффициентов многочлена.

Однозначная разрешимость следует из хорошо известного факта, что определитель этой системы (*опредетитель Вандермонда*)

$$\begin{vmatrix} 1 & x_0 & \dots & x_0^n \\ 1 & x_1 & \dots & x_1^n \\ \dots & \dots & \dots & \dots \\ 1 & x_n & \dots & x_n^n \end{vmatrix} = \prod_{0 \leq j < i \leq n} (x_i - x_j) \neq 0.$$

Замечание 1. На практике система (4.2) не используется для вычисления коэффициентов интерполяционного многочлена, т. к. она является плохо обусловленной.

Будем искать *полином Лагранжа* в виде

$$L_n(x) = l_0(x) + l_1(x) + \dots + l_n(x),$$

где $l_i(x)$ – многочлен степени n , причем

$$l_i(x_k) = \begin{cases} y_i, & \text{если } i = k, \\ 0, & \text{если } i \neq k. \end{cases} \quad (4.3)$$

Для выполнения последнего условия многочлены $l_i(x)$ составим следующим способом:

$$l_i(x) = c_i(x - x_0)(x - x_1)\dots(x - x_{i-1})(x - x_{i+1})\dots(x - x_n).$$

Здесь c_i – постоянный коэффициент, значение которого находится из (4.3):

$$c_i = \frac{y_i}{(x_i - x_0)(x_i - x_1)\dots(x_i - x_{i-1})(x_i - x_{i+1})\dots(x_i - x_n)}.$$

Для полинома Лагранжа получили окончательную формулу:

$$L_n(x) = \sum_{i=0}^n y_i \frac{(x - x_0)(x - x_1)\dots(x - x_{i-1})(x - x_{i+1})\dots(x - x_n)}{(x_i - x_0)(x_i - x_1)\dots(x_i - x_{i-1})(x_i - x_{i+1})\dots(x_i - x_n)}. \quad (4.4)$$

В частности полиномы Лагранжа первой и второй степени соответственно имеют вид

$$L_1(x) = y_0 \frac{x_1 - x}{h} + y_1 \frac{x - x_0}{h},$$

$$L_2(x) = y_0 \frac{(x-x_1)(x-x_2)}{2h^2} + y_1 \frac{(x-x_0)(x-x_2)}{(-h^2)} + y_2 \frac{(x-x_0)(x-x_1)}{2h^2}.$$

Приведем без доказательства теорему о погрешности интерполяции.

Теорема. Пусть функция f дифференцируема $n+1$ раз на отрезке $[a, b]$, содержащим узлы интерполяции $x_i, i=0, 1, \dots, n$. Тогда для погрешности интерполяции в каждой точке отрезка $[a, b]$ справедливо равенство

$$f(x) - P_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \omega_{n+1}(x), \quad (4.5)$$

в котором $\omega_{n+1}(x) = (x-x_0)(x-x_1)\dots(x-x_n)$, а $\xi \in (a, b)$.

Следствие. Пусть $M_{n+1} = \max_{[a,b]} |f^{(n+1)}(x)|$, тогда справедлива оценка максимума модуля интерполяции на отрезке $[a, b]$, имеющая вид

$$\max_{[a,b]} |f(x) - P_n(x)| \leq \frac{M_{n+1}}{(n+1)!} \max_{[a,b]} |\omega_{n+1}(x)|. \quad (4.6)$$

Характерное поведение функции $\omega_{n+1}(x) = (x-x_0)(x-x_1)\dots(x-x_n)$ представлено на рис. 17. Внутри отрезка $[x_0, x_n]$ множитель $\omega_{n+1}(x)$ не обесценивает оценку (4.6). Случай, когда x выходит за пределы отрезка $[x_0, x_n]$, называют *экстраполяцией* функции f . Если аргумент x выходит за пределы отрезка $[x_0, x_n]$, то значение функции $\omega_{n+1}(x)$ резко возрастает, что показывает ненадежность *экстраполяции* функции $f(x)$ для x , удаленных от концов отрезка $[x_0, x_n]$.

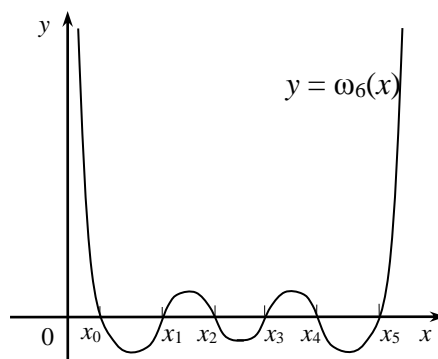


Рис. 17

4.2. Интерполяция с кратными узлами. Полином Эрмита

Пусть теперь требуется по известным значениям функции и ее производных

$$y_i = f(x_i), \quad y'_i = f'(x_i), \dots, \quad y_i^{(k_i-1)} = f^{(k_i-1)}(x_i), \quad i = 0, 1, \dots, m$$

«восстановить» исходную функцию $y = f(x)$. В этом случае число k_i называют кратностью корня x_i . В каждой точке x_i задано k_i величин. Пусть $n = k_0 + k_1 + \dots + k_m - 1$, можно доказать, что существует единственный многочлен порядка n , удовлетворяющий условиям:

$$P_n(x_i) = y_i, \quad P'_n(x_i) = y'_i, \dots, \quad P_n^{(k_i-1)}(x_i) = y_i^{(k_i-1)}, \quad i = 0, 1, \dots, m \quad (4.7)$$

Этот многочлен называют интерполяционным *многочленом с кратными узлами или интерполяционным многочленом Эрмита*. В общем случае построение многочлена Эрмита приводит к довольно громоздким выражениям. Поэтому ограничимся двумя простыми случаями.

1. Пусть на концах отрезка $[x_0, x_1]$ заданы значения функции и ее первой производной y_0, y'_0, y_1, y'_1 . *Кубический интерполяционный многочлен Эрмита* $P_3(x)$, удовлетворяющий условиям

$$P_3(x_0) = y_0, \quad P_3(x_1) = y_1, \quad P'_3(x_0) = y'_0, \quad P'_3(x_1) = y'_1,$$

может быть представлен в следующей форме:

$$P_3(x) = y_0 \frac{(x_1 - x)^2 (2(x - x_0) + h)}{h^3} + y'_0 \frac{(x_1 - x)^2 (x - x_0)}{h^2} + y_1 \frac{(x - x_0)^2 (2(x_1 - x) + h)}{h^3} + y'_1 \frac{(x - x_0)^2 (x - x_1)}{h^2}, \quad (4.8)$$

где $h = x_1 - x_0$.

2. Пусть в точке x_0 задано n значений $y_0, y'_0, y''_0, \dots, y_0^{(n-1)}$. Тогда, как следует из формулы Тейлора, многочлен $P_n(x)$ представляется в виде

$$P_n(x) = \sum_{k=0}^{n-1} y_0^{(k)} \frac{(x - x_0)^k}{k!}. \quad (4.9)$$

Замечание 2. Для погрешности интерполяции с кратными узлами справедливы формулы, аналогичные (4.5), (4.6), в которых

$$\omega_{n+1}(x) = (x - x_0)^{k_0} (x - x_1)^{k_1} \dots (x - x_m)^{k_m}.$$

4.3. Конечные разности. Интерполяционный полином Ньютона для равноотстоящих узлов

Для функций, заданных (4.1) с постоянным шагом, определим разности между значениями функции в соседних узлах интерполяции:

$$\Delta y_i = y_{i+1} - y_i, \quad (4.10)$$

$$\Delta^2 y_i = \Delta y_{i+1} - \Delta y_i = (y_{i+2} - y_{i+1}) - (y_{i+1} - y_i) = y_{i+2} - 2y_{i+1} + y_i, \quad (4.11)$$

$$\begin{aligned} \Delta^3 y_i &= \Delta^2 y_{i+1} - \Delta^2 y_i = \\ &= (y_{i+3} - 2y_{i+2} + y_{i+1}) - (y_{i+2} - 2y_{i+1} + y_i) = \\ &= y_{i+3} - 3y_{i+2} + 3y_{i+1} - y_i, \end{aligned} \quad (4.12)$$

которые соответственно называются конечными разностями первого, второго и третьего порядка.

Методом математической индукции можно доказать, что

$$\Delta^k y_i = y_{i+k} - ky_{i+k-1} + \frac{k(k-1)}{2!} y_{i+k-2} - \dots + (-1)^k y_i. \quad (4.13)$$

Будем искать интерполяционный многочлен Ньютона в виде:

$$\begin{aligned} P_n(x) &= a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \dots + \\ &+ a_n(x - x_0)(x - x_1)\dots(x - x_{n-1}). \end{aligned} \quad (4.14)$$

Значения коэффициентов a_i найдем из условий (4.2). Полагая в (4.14) $x = x_0$, найдем $P_n(x_0) = a_0 = y_0$. Далее, полагая $x = x_1$, $x = x_2$, получаем

$$P_n(x_1) = a_0 + a_1(x_1 - x_0) = y_0 + a_1 h = y_1,$$

откуда $a_1 = \frac{\Delta y_0}{h}$;

$$P_n(x_2) = a_0 + a_1(x_2 - x_0) + a_2(x_2 - x_0)(x_1 - x_0) = y_0 + \frac{\Delta y_0}{h} 2h + a_2 2h^2 = y_2,$$

или

$$2h^2 a_2 = y_2 - 2\Delta y_0 - y_0 = y_2 - 2(y_1 - y_0) - y_0 = y_2 - 2y_1 + y_0 = \Delta^2 y_0,$$

откуда $a_2 = \frac{\Delta^2 y_0}{2!h^2}$.

Затем, проведя аналогичные выкладки, можно получить $a_3 = \frac{\Delta^3 y_0}{3!h^3}$.

В общем случае выражение для a_k будет иметь вид

$$a_k = \frac{\Delta^k y_0}{k!h^k}.$$

Первая интерполяционная формула Ньютона записывается в форме

$$P_n(x) = y_0 + \frac{\Delta y_0}{h}(x - x_0) + \frac{\Delta^2 y_0}{2!h^2}(x - x_0)(x - x_1) + \dots + \frac{\Delta^n y_0}{n!h^n}(x - x_0)(x - x_1)\dots(x - x_{n-1}).$$

Когда значение аргумента находится ближе к концу отрезка интерполяции, используется *вторая интерполяционная формула Ньютона*, которая получается, если искать интерполяционный многочлен в виде:

$$P_n(x) = a_0 + a_1(x - x_n) + a_2(x - x_n)(x - x_{n-1}) + \dots + a_n(x - x_n)(x - x_{n-1})\dots(x - x_1).$$

Выражение для коэффициентов a_k будет иметь вид

$$a_k = \frac{\Delta^k y_{n-k}}{k!h^k}. \quad (4.15)$$

В практическом плане формула (4.14) обладает преимуществом перед формулой Лагранжа (4.4). Если нужно добавить в таблицу еще один узел x_{n+1} , то формулу Лагранжа (4.4) придется перестраивать заново, в то время как, в формулы Ньютона добавляется лишь одно очередное слагаемое.

Замечание 3. Для достаточно гладких функций $f(x)$ величину

$$\varepsilon_n = |P_{n+1}(x) - P_n(x)|$$

можно использовать для практической оценки погрешности интерполяции по формулам Ньютона.

4.4. Минимизация оценки погрешности интерполяции. Многочлены Чебышева

Предположим, что вычисление значений функции $f(x)$ – достаточно трудоемкая операция, и целесообразнее заменить прямое вычисление функции f вычислением ее интерполяционного многочлена P_n . Для такой замены необходимо один раз получить таблицу значений функции f в выбранных точках x_0, x_1, \dots, x_n отрезка $[a, b]$. При этом выбор узлов интерполяции должен быть таким, чтобы погрешность интерполяции была наименьшей.

Из теоремы (4.1) для верхней границы погрешность интерполяции имеем

$$\bar{\Delta}(P_n) = \max_{[a,b]} |f(x) - P_n(x)| = \frac{M_{n+1}}{(n+1)!} \max_{[a,b]} |\omega_{n+1}(x)|. \quad (4.16)$$

Поставим задачу определения набора узлов x_0, x_1, \dots, x_n , при котором величина $\bar{\Delta}(P_n)$ минимальна. Решение этой задачи опирается на свойства многочленов Чебышева.

Многочлены Чебышева определяются рекуррентным способом:

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x), \quad n \geq 2. \quad (4.17)$$

Явные формулы при $n = 2, 3$:

$$\begin{aligned} T_2(x) &= 2xT_1(x) - T_0(x) = 2x^2 - 1, \\ T_3(x) &= 2xT_2(x) - T_1(x) = 4x^3 - 3x. \end{aligned}$$

Приведем основные свойства многочленов Чебышева.

1. При четном n многочлен $T_n(x)$ является четной функцией, при нечетном n многочлен $T_n(x)$ – нечетной функцией.

2. При $n \geq 1$ старший коэффициент многочлена $T_n(x)$ равен 2^{n-1} .

3. Для $x \in [-1, 1]$ имеется явная формула для многочлена $T_n(x)$:

$$T_n(x) = \cos(n \arccos x). \quad (4.18)$$

4. При $n \geq 1$ многочлен $T_n(x)$ имеет ровно n действительных корней $x_i \in [-1, 1]$, вычисляемых по формуле

$$x_k = \cos \frac{(2k+1)\pi}{2n}, \quad k = 0, 1, \dots, n-1. \quad (4.19)$$

5. При $n \geq 0$ $\max_{[-1,1]} |T_n(x)| = 1$. Если $n \geq 1$, то этот максимум достигается ровно в $n + 1$ точках, которые находятся по формуле

$$x_k = \cos \frac{k\pi}{n}.$$

При этом $T_n(x_k) = (-1)^k$, т. е. максимумы и минимумы многочлена Чебышева чередуются.

6. Среди всех многочленов фиксированной степени $n \geq 1$ со старшим единичным коэффициентом, наименьшее уклонение от нуля (равное 2^{1-n}) имеет многочлен

$$\bar{T}_n(x) = 2^{1-n} T_n(x), \text{ т. е.}$$

$$2^{1-n} = \max_{[-1,1]} |\bar{T}_n(x)| \leq \max_{[-1,1]} |P_n(x)|.$$

Геометрически корни уравнения и точки экстремума многочлена $T_n(x)$ можно получить, если полуокружность единичного радиуса разделить на $2n$ равных частей, а затем спроецировать на ось абсцисс.

Пронумеровав полученные точки вдоль оси абсцисс справа налево, получим корни уравнения – точки с нечетными номерами (на рис. 18 они помечены кружочками), а точки экстремума – с четными номерами (на рис. 18 они помечены крестиками). Корни и точки экстремума сгущаются к концам интервала.

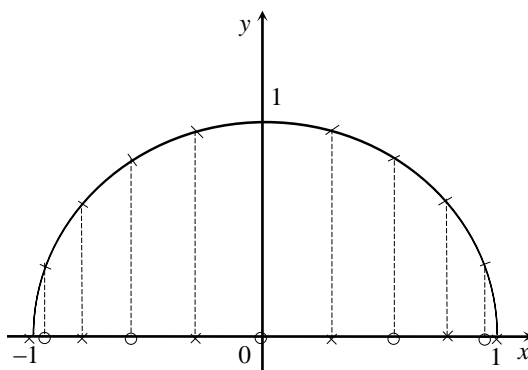


Рис. 18

Пусть отрезок интерполяции $[-1,1]$, в этом случае оценка (4.16) будет минимальна, если минимальна величина $\max_{[-1,1]} |\omega_{n+1}(x)|$.

В силу свойств 4...6 многочленов Чебышева минимальное уклонение многочлена

$$\omega_{n+1}(x) = (x - x_0)(x - x_1)\dots(x - x_n)$$

дает набор узлов

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

т. к. в этом случае $\omega_{n+1}(x) = \bar{T}_{n+1}(x)$.

Заметим, что при этом выборе $\bar{\Delta}(P_n) = \frac{M_{n+1}}{(n+1)!2^n}$.

Пусть теперь отрезок интерполяции $[a, b]$ произволен.

Заменой

$$x = \frac{a+b}{2} + t \frac{b-a}{2}, \quad t \in [-1, 1]$$

он приводится к стандартному отрезку $[-1, 1]$. Легко увидеть, что в этом случае

$$\omega_{n+1}(x) = \left(\frac{b-a}{2}\right)^{n+1} (t-t_0)(t-t_1)\dots(t-t_n),$$

$$\bar{\Delta}(P_n) = \frac{M_{n+1}}{(n+1)!2^n} \left(\frac{b-a}{2}\right)^{n+1}.$$

4.5. Интерполяция сплайнами

Сплайн – это функция, которая на каждом частичном отрезке интерполяции является алгебраическим многочленом невысокого порядка, а на всем заданном отрезке непрерывна с несколькими своими производными.

На каждом частичном отрезке $[x_{i-1}, x_i]$ с длиной $h_i = x_i - x_{i-1}$ будем искать кубический сплайн в виде

$$S(x) = a_i + b_i(x - x_i) + \frac{c_i}{2}(x - x_i)^2 + \frac{d_i}{6}(x - x_i)^3, \quad i = 0, 1, \dots, n; \quad (4.20)$$

с неизвестной четверкой коэффициентов a_i, b_i, c_i, d_i .

При этом очевидно, что

$$S'(x) = b_i + c_i(x - x_i) + \frac{d_i}{2}(x - x_i)^2, \quad (4.21)$$

$$S''(x) = c_i + d_i(x - x_i). \quad (4.22)$$

Потребуем совпадения сплайна $S(x)$ в узлах с табличными значениями (4.1):

$$S(x_i) = a_i = y_i, \quad (4.23)$$

$$S(x_{i-1}) = a_i - b_i h_i + \frac{c_i}{2} h_i^2 - \frac{d_i}{6} h_i^3 = y_{i-1}. \quad (4.24)$$

Требуя непрерывности сплайна, а также непрерывности первой и второй производной, во всех точках, включая внутренние узлы, из равенств (4.21)–(4.24) имеем

$$b_i h_i - \frac{c_i}{2} h_i^2 + \frac{d_i}{6} h_i^3 = a_i - y_{i-1}, \quad i = 0, 1, \dots, n;$$

$$S'(x_{i-1} + 0) = S'(x_{i-1} - 0), \quad \text{или} \quad b_i - c_i h_i + \frac{d_i}{2} h_i^2 = b_{i-1};$$

$$S''(x_{i-1} + 0) = S''(x_{i-1} - 0), \quad \text{или} \quad c_i - d_i h_i = c_{i-1}.$$

Если дополнительно потребовать для *естественного* кубического сплайна нулевой кривизны на концах отрезка, (т. е. равенства нулю второй производной), то получим еще два соотношения:

$$c_n = 0, \\ , \quad c_1 - d_1 h_1 = 0.$$

Исключив a_i , по формуле (4.23), для определения b_i, c_i, d_i , $i = 1, 2, \dots, n$ получаем систему из $3n$ уравнений:

$$b_i h_i - \frac{c_i}{2} h_i^2 + \frac{d_i}{6} h_i^3 = y_i - y_{i-1},$$

$$c_i h_i - \frac{d_i}{2} h_i^2 = b_i - b_{i-1},$$

$$d_i h_i = c_i - c_{i-1},$$

$$c_n = 0,$$

$$c_1 - d_1 h_1 = 0.$$

Если последовательно исключить d_i из третьего уравнения системы, а затем b_i из первого уравнения и подставить полученные выражения во второе уравнение, то получим линейную систему с трехдиагональной матрицей относительно c_i .

$$h_i c_{i-1} + 2(h_i + h_{i+1})c_i + h_{i+1}c_{i+1} = 6 \left(\frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i} \right),$$

$$i = 0, 1, \dots, n-1. \quad (4.25)$$

Кроме того, $c_n = 0$, и для единообразия положили $c_0 = 0$. Система (4.25) имеет диагональное преобладание, и решение таких систем эффективно осуществляется методом прогонки.

Приведем без доказательства теорему о погрешности интерполяции кубическим сплайном.

Теорема. Пусть функция дифференцируема четыре раза на отрезке $[a, b]$, дополнительно удовлетворяет условию $f''(a) = f''(b)$. Тогда для погрешности интерполяции функции f и ее производных в каждой точке отрезка $[a, b]$ имеют место оценки:

$$\begin{aligned} |f(x) - S(x)| &\leq M_4 h^4, \\ |f'(x) - S'(x)| &\leq M_4 h^3, \\ |f''(x) - S''(x)| &\leq M_4 h^2, \end{aligned}$$

в которых $M_4 = \max_{[a,b]} |f^{(4)}(x)|$, $h = \max_i h_i$.

4.6. Метод наименьших квадратов

Если число узлов x_0, x_1, \dots, x_n , в которых производятся измерения, велико, а значения y_i в точках x_i определяются приближенно в связи с неизбежными ошибками измерения, то строить функцию $F(x)$, в виде суммы большого количества слагаемых и добиваться ее точного равенства в узлах сетки значениям y_i становится нецелесообразным.

В методе наименьших квадратов аппроксимирующая функция $F(x)$ ищется в виде суммы, содержащей сравнительно небольшое число слагаемых:

$$F(x) = \sum_{k=0}^m a_k \varphi_k(x), \quad m < n. \quad (4.26)$$

Здесь $\varphi_0(x), \varphi_1(x), \dots, \varphi_m(x)$ – заданные базисные, линейно независимые функции; a_0, a_1, \dots, a_m – параметры модели, являющиеся коэффициентами обобщенного многочлена $F(x)$.

Пусть $\bar{y}_i = F(x_i)$. Составим суммарную квадратичную погрешность приближения:

$$J = \sum_{i=0}^n (y_i - \bar{y}_i)^2 = \sum_{i=0}^n (y_i - \sum_{k=0}^m a_k \varphi_k(x_i))^2. \quad (4.27)$$

Возникает задача найти такой набор коэффициентов a_0, a_1, \dots, a_m , при котором суммарная квадратичная погрешность (4.27) оказывается минимальной. При этом $F(x)$ называют наилучшим приближением по методу наименьших квадратов.

Построение наилучшего приближения сводится к классической задаче математического анализа об экстремуме функции нескольких переменных. Необходимым условием экстремума является выполнение в точках экстремума равенств $\frac{\partial}{\partial a_l} J = 0$:

$$\frac{\partial}{\partial a_l} J = -2 \sum_{i=0}^n \varphi_l(x_i) (y_i - \sum_{k=0}^m a_k \varphi_k(x_i)) = 0, \quad l = 0, 1, \dots, m. \quad (4.28)$$

Меняя порядок суммирования, уравнение (4.28) можно записать в виде нормальной системы линейных уравнений относительно a_0, a_1, \dots, a_m :

$$\sum_{k=0}^m d_{lk} a_k = b_l, \quad l = 0, 1, \dots, m, \quad (4.29)$$

где

$$d_{lk} = \sum_{i=0}^n \varphi_l(x_i) \varphi_k(x_i), \quad b_l = \sum_{i=0}^n \varphi_l(x_i) y_i.$$

Представим систему (4.29) для определения коэффициентов a_0, a_1 для случая, когда приближающая функция $F(x) = a_0 \varphi_0(x) + a_1 \varphi_1(x)$ есть линейная комбинация двух линейно независимых известных функций $\varphi_0(x), \varphi_1(x)$ подробно:

$$\begin{pmatrix} \sum_{i=0}^n \varphi_0^2(x_i) & \sum_{i=0}^n \varphi_0(x_i) \varphi_1(x_i) \\ \sum_{i=0}^n \varphi_0(x_i) \varphi_1(x_i) & \sum_{i=0}^n \varphi_1^2(x_i) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^n \varphi_0(x_i) y_i \\ \sum_{i=0}^n \varphi_1(x_i) y_i \end{pmatrix}. \quad (4.30)$$

Матрица коэффициентов d_{lk} является диагональной и положительно определенной, если $\varphi_0(x), \varphi_1(x), \dots, \varphi_m(x)$ – линейно независимые функции в точках x_0, x_1, \dots, x_n .

Пусть система (4.29) имеет единственное решение a_0, a_1, \dots, a_m . Рассмотрим любой другой набор переменных $\bar{a}_0, \bar{a}_1, \dots, \bar{a}_m$:

$$\bar{a}_i = a_i + \Delta a_i, \quad i = 0, 1, \dots, m.$$

Тогда можно показать, что имеет место неравенство

$$J(\bar{a}_0, \bar{a}_1, \dots, \bar{a}_m) = J(a_0, a_1, \dots, a_m) + \sum_{i=0}^n \left(\sum_{k=0}^m a_k \varphi_k(x_i) \right)^2 \geq J(a_0, a_1, \dots, a_m).$$

Последнее неравенство показывает, что решение системы (4.29) действительно минимизирует суммарную погрешность (4.27).

Замечание 4. Будучи формально линейно независимой, система $\varphi_0(x), \varphi_1(x), \dots, \varphi_m(x)$ функций может оказаться очень близкой к линейно зависимой системе, что делает задачу плохо обусловленной. Нормальная система (4.29) оказывается, как правило, плохо обусловленной уже для набора степенных функций $1, x, \dots, x^5$. Чтобы избежать этого, иногда предварительно проводят ортогонализацию системы функции $\varphi_0(x), \varphi_1(x), \dots, \varphi_m(x)$.

Пример 1. Найти приближающую функцию в виде линейной функции $F(x) = ax + b$.

Здесь $\varphi_0(x) = 1, \varphi_1(x) = x$; неизвестные коэффициенты – a и b . Составим суммарную квадратичную погрешность приближения:

$$J(a, b) = \sum_{i=0}^n (y_i - (ax_i + b))^2.$$

Из равенств $\frac{\partial}{\partial a} J = 0, \frac{\partial}{\partial b} J = 0$ следует:

$$\sum_{i=0}^n (y_i - ax_i - b)x_i = 0, \quad \sum_{i=0}^n (y_i - ax_i - b) = 0.$$

Последние равенства можно переписать в виде

$$\left(\frac{1}{n+1} \sum_{i=0}^n x_i^2 \right) a + \left(\frac{1}{n+1} \sum_{i=0}^n x_i \right) b = \frac{1}{n+1} \sum_{i=0}^n y_i x_i,$$

$$\left(\frac{1}{n+1} \sum_{i=0}^n x_i \right) a + b = \frac{1}{n+1} \sum_{i=0}^n y_i. \quad (4.31)$$

Из системы (4.31) находим a и b .

Пример 2. Найти приближающую функцию в виде квадратичного трехчлена

$$F(x) = ax^2 + bx + c.$$

Здесь $\varphi_0(x) = 1$, $\varphi_2(x) = x$; $\varphi_3(x) = x^2$; неизвестные коэффициенты — a , b и c . Составим суммарную квадратичную погрешность приближения:

$$J(a, b) = \sum_{i=0}^n (y_i - (ax_i^2 + bx_i + c))^2.$$

Из равенств $\frac{\partial}{\partial a} J = 0$, $\frac{\partial}{\partial b} J = 0$ и $\frac{\partial}{\partial c} J = 0$ следует:

$$\sum_{i=0}^n (y_i - ax_i^2 - bx_i - c)x_i^2 = 0, \quad \sum_{i=0}^n (y_i - ax_i^2 - bx_i - c)x_i = 0,$$

$$\sum_{i=0}^n (y_i - ax_i^2 - bx_i - c) = 0.$$

Последние равенства можно переписать в следующей форме:

$$\begin{aligned} \left(\frac{1}{n+1} \sum_{i=0}^n x_i^4 \right) a + \left(\frac{1}{n+1} \sum_{i=0}^n x_i^3 \right) b + \left(\frac{1}{n+1} \sum_{i=0}^n x_i^2 \right) c &= \frac{1}{n+1} \sum_{i=0}^n y_i x_i^2, \\ \left(\frac{1}{n+1} \sum_{i=0}^n x_i^3 \right) a + \left(\frac{1}{n+1} \sum_{i=0}^n x_i^2 \right) b + \left(\frac{1}{n+1} \sum_{i=0}^n x_i \right) c &= \frac{1}{n+1} \sum_{i=0}^n y_i x_i, \\ \left(\frac{1}{n+1} \sum_{i=0}^n x_i^2 \right) a + \left(\frac{1}{n+1} \sum_{i=0}^n x_i \right) b + c &= \frac{1}{n+1} \sum_{i=0}^n y_i. \end{aligned} \quad (4.32)$$

Из последней системы находим a , b и c .

Пример 3. Найти приближающую функцию в виде степенной функции $F(x) = ax^m$ с параметрами $a > 0$ и m .

Прологарифмируем данную функцию в предположение, что все ее значения положительны:

$$\ln F(x) = \ln a + m \ln x.$$

Функция $\ln F(x)$ будет приближающей для функции $\ln f(x)$. Введем новую переменную $u = \ln x$, тогда $\ln F(x) = \Phi(u) = \ln a + tu$, т. е. функция $\Phi(u)$ будет линейной относительно $u = \ln x$. Пришли к линейной задаче определения новых параметров $\ln a$ и t по значениям $\ln f_i$, рассмотренной в прим. 1.

Пример 4. Найти приближающую функцию в виде показательной функции $F(x) = ae^{mx}$ с параметрами $a > 0$ и m .

Прологарифмируем данную функцию в предположение, что все ее значения положительны:

$$\ln F(x) = \ln a + mx.$$

Функция $\ln F(x)$ будет приближающей для функции $\ln f(x)$. Пришли к линейной задаче определения новых параметров $\ln a$ и m по значениям $\ln f_i$, рассмотренной в прим. 1.

Пример 5. Найти приближающую функцию в виде дробно-линейной функции $F(x) = \frac{1}{ax + b}$ с параметрами a и b .

Обратим данную функцию в предположение, что все ее значения отличны от нуля.

$$\frac{1}{F(x)} = ax + b.$$

Функция $1/F(x)$ будет приближающей для функции $1/f(x)$. Пришли вновь к линейной задаче определения параметров a и b по значениям $1/f_i$, рассмотренной в прим. 1.

Замечание 5. Задачи определения приближающих функций вида $F(x) = \frac{x}{ax + b}$, $F(x) = \frac{a}{x} + b$, $F(x) = a \ln x + b$ так же приводятся к линейной задаче, рассмотренной в прим. 1.

5. ЧИСЛЕННОЕ ИНТЕГРИРОВАНИЕ

5.1. Простейшие квадратурные формулы.

Формула прямоугольников, трапеции и формула Симпсона

Для вычисления значения определенного интеграла на практике широко используют квадратурные формулы – приближенные равенства вида

$$\int_a^b f(x)dx \approx \sum_{i=0}^n A_i f(\xi_i). \quad (5.1)$$

Здесь ξ_i – некоторые точки из отрезка $[a, b]$ – узлы квадратурной формулы; A_i – числовые коэффициенты, называемые *весами квадратурной формулы*.

Сумма $\sum_{i=0}^n A_i f(\xi_i)$, принимаемая за приближенное значение интеграла, называется *квадратурной суммой*.

Если для любого многочлена $P_k(x)$ степени не выше m квадратурная формула (5.1) дает точное значение интеграла

$$\int_a^b P_k(x)dx = \sum_{i=0}^n A_i P_k(\xi_i),$$

то говорят, что квадратурная формула *точна для многочленов степени m* .

Рассмотрим простейшие квадратурные формулы.

Для этого разобьем отрезок $[a, b]$ на элементарные отрезки $[x_{i-1}, x_i]$ точками $a = x_0 < x_1 < \dots < x_n = b$.

Интеграл при этом

$$I = \int_a^b f(x)dx = \sum_{i=1}^n I_i, \quad \text{где} \quad I_i = \int_{x_{i-1}}^{x_i} f(x)dx.$$

Для краткости введем обозначения:

$$f_i = f(x_i), \quad f_{i-1/2} = f(x_{i-1/2}),$$

где $x_{i-1/2} = (x_{i-1} + x_i) / 2$ – середина элементарного отрезка.

Пусть шаг $h_i = x_i - x_{i-1}$ будет постоянным.

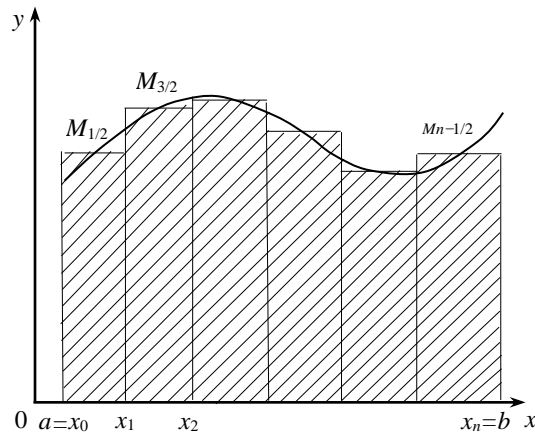


Рис. 19

Приближим на каждом элементарном отрезке площадь криволинейной трапеции площадью прямоугольника, с основанием $h = x_i - x_{i-1}$ и высотой $f_{i-1/2}$ (рис. 19). Площадь каждого элементарного прямоугольника $I_i \approx hf_{i-1/2}$, тогда просуммировав по всем элементарным отрезкам, для площади криволинейной трапеции имеем *квадратурную формулу прямоугольников*:

$$I = \sum_{i=1}^n I_i \approx I_{np}^h = h \sum_{i=1}^n f_{i-1/2}. \quad (5.2)$$

Здесь все A_i – веса квадратурной формулы (5.2) равны h .

Иногда используют формулы

$$I_{np} = h \sum_{i=0}^{n-1} f_i, \quad I_{np} = h \sum_{i=1}^n f_i, \quad (5.3)$$

называемые соответственно *составными квадратурными формулами левых и правых прямоугольников* (рис. 20, а и б).

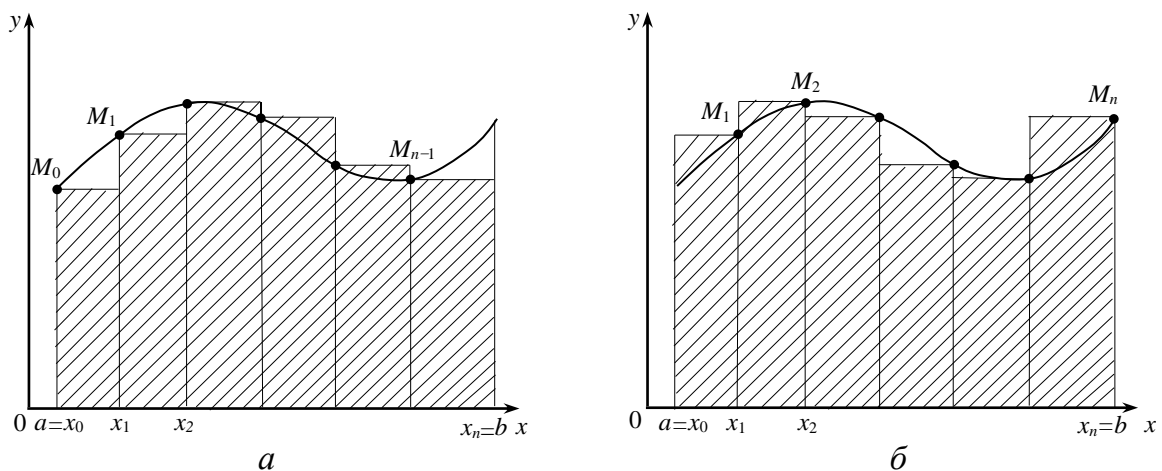


Рис. 20

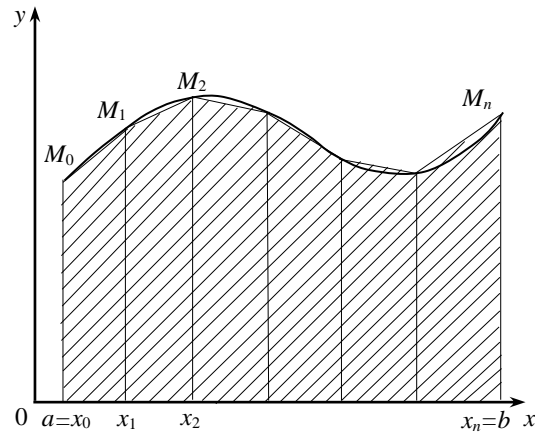


Рис. 21

Приближим на каждом элементарном отрезке площадь криволинейной трапеции площадью трапеции, с высотой $h = x_i - x_{i-1}$ и основаниями f_i и f_{i-1} , соединив точки $M_{i-1}(x_{i-1}, f_{i-1})$ и $M_i(x_i, f_i)$ (рис. 21).

Площадь каждой элементарной криволинейной трапеции $I_i \approx h(f_{i-1} + f_i)/2$, тогда просуммировав по всем элементарным отрезкам, для площади криволинейной трапеции имеем *квадратурную формулу трапеций*:

$$I \approx I_{mp}^h = h \left[\frac{f_0}{2} + f_1 + f_2 + \dots + f_{n-1} + \frac{f_n}{2} \right] = h \left[\frac{f_0 + f_n}{2} + \sum_{i=1}^{n-1} f_i \right]. \quad (5.4)$$

Здесь A_i – веса квадратурной формулы (5.4) равны h , за исключением $A_0 = A_n = h/2$.

Вывод формулы Симпсона развивает описанный подход дальше. Но теперь для аппроксимации функции используется не кусочно-линейное, а кусочно-квадратичное интерполирование.

Будем считать n четным и сгруппируем отрезки парами: $[a, x_1]$, $[x_1, x_2]$; вторая пара $[x_2, x_3]$, $[x_3, x_4]$ и т. д. Для каждого двойного отрезка $[x_{2j-2}, x_{2j}]$ построим интерполяционный полином второй степени в форме Лагранжа, принимающий в узлах значения $f_{2j-2} = f(x_{2j-2})$, $f_{2j-1} = f(x_{2j-1})$, $f_{2j} = f(x_{2j})$.

$$f(x) \approx L_2(x) = f(x_{2j-2}) \frac{(x - x_{2j-1})(x - x_{2j})}{2h^2} + f(x_{2j-1}) \frac{(x - x_{2j-2})(x - x_{2j})}{(-h^2)} + f(x_{2j}) \frac{(x - x_{2j-2})(x - x_{2j-1})}{2h^2}, \quad x \in [x_{2j-2}, x_{2j}], \quad 1 \leq j \leq n/2. \quad (5.5)$$

Проинтегрировав (5.5) по отрезку $[x_{2j-2}, x_{2j}]$, получим

$$I_j \approx \int_{x_{2j-2}}^{x_{2j}} L_2(x) dx = h/3 \{f(x_{2j-2}) + 4f(x_{2j-1}) + f(x_{2j})\}, \quad h = \frac{b-a}{n}.$$

Тогда *квадратурная формула Симпсона* или *парабол* примет вид

$$\begin{aligned} I &= \sum_{j=1}^{n/2} I_j \approx I_{Сум}^h = \frac{h}{3} \sum_{j=1}^{n/2} \{f(x_{2j-2}) + 4f(x_{2j-1}) + f(x_{2j})\} = \\ &= \frac{b-a}{3n} \{f(a) + 4f(x_1) + 2f(x_2) + 4f(x_3) + 2f(x_4) + \dots + \\ &\quad + 2f(x_{n-2}) + 4f(x_{n-1}) + f(b)\}. \end{aligned} \quad (5.6)$$

Напомним, что n – четное число.

5.2. Оценка погрешности квадратурных формул прямоугольников, трапеции и формулы Симпсона

Докажем теорему об оценке погрешности формул прямоугольников и трапеций. Предположим, что подынтегральная функция f достаточно гладкая и $M_k = \max_{[a,b]} |f^{(k)}(x)|$.

Теорема Пусть функция f дважды непрерывно дифференцируема на $[a,b]$. Тогда для составных квадратурных формул прямоугольников и трапеций справедливы следующие оценки погрешности:

$$|I - I_{np}^h| \leq \frac{M_2(b-a)}{24} h^2, \quad (5.7)$$

$$|I - I_{mp}^h| \leq \frac{M_2(b-a)}{12} h^2. \quad (5.8)$$

Доказательство. Используя формулу Тейлора

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

где $x \in [x_{i-1}, x_i]$, $\xi \in (x_{i-1}, x_i)$, имеем

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

$$|R_i| \leq \frac{M_2}{2} \int_{x_{i-1}}^{x_i} (x - x_{i-1/2})^2 dx = \frac{M_2}{6} (x_i - x_{i-1/2})^3 = \frac{M_2}{24} h^3.$$

Так как $R = \sum_{i=1}^n R_i$, то $|R| \leq \frac{M_2}{24} h^3 n = \frac{M_2(b-a)}{24} h^2$ ($h_n = b - a$).

Для вывода оценки (5.8) заметим, что отрезок, соединяющий точки M_{i-1} и M_i (рис. 21) представляет собой график интерполяционного многочлена первой степени $L_1(x) = f_{i-1} \frac{x_i - x}{h} + f_i \frac{x - x_{i-1}}{h}$. Поэтому верно равенство

$$R_i = \int_{x_{i-1}}^{x_i} f(x) dx - \frac{h}{2} (f(x_{i-1}) + f(x_i)) = \int_{x_{i-1}}^{x_i} (f(x) - L_1(x)) dx.$$

Используя теорему (6.1) об оценке погрешности линейной интерполяции и следствие к ней, имеем

$$|R_i| \leq \frac{M_2}{2} \int_{x_{i-1}}^{x_i} (x - x_{i-1})(x_i - x) dx = \frac{M_2}{12} h^3.$$

Следовательно, для $R = |I - I_{\text{тр}}^h|$ справедлива оценка

$$|R| \leq \sum_{i=1}^n |R_i| \leq \frac{M_2}{12} h^3 n = \frac{M_2(b-a)}{12} h^2.$$

Аналогично доказывается теорема об оценке погрешности формулы Симпсона.

Теорема. Пусть функция $f(x)$ имеет непрерывную производную четвертого порядка. Тогда для формулы Симпсона справедлива оценка погрешности:

$$|I - I_{\text{Сим}}^h| \leq \frac{M_4(b-a)}{2880} h^4. \quad (5.9)$$

Замечания:

1. Оценки (5.7)–(5.9) означают, что формулы прямоугольников и трапеций имеют второй порядок точности относительно h , а формула Симпсона – четвертый порядок точности.

2. Так как для многочленов первой степени и для многочленов третьей степени $M_2 = 0$ и $M_4 = 0$, то квадратурная формула прямоугольников и трапеции точна для многочленов первой степени, а формула Симпсона – для многочленов третьей степени.

5.3. Квадратурные формулы интерполяционного типа

Интеграл $I = \int_a^b f(x)dx$ представляют в виде суммы интегралов по элементарным отрезкам

$$I = \int_a^b f(x)dx = \sum_{i=1}^n I_i, \text{ где } I_i = \int_{x_{i-1}}^{x_i} f(x)dx.$$

На каждом отрезке $[x_{i-1}, x_i]$ функцию f аппроксимируют легко интегрируемой функцией $g_i(x)$. В результате получается составная формула

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

Рассмотрим случай, когда в качестве функции g_i на отрезке $[x_{i-1}, x_i]$ берутся интерполяционные многочлены степени m . Пусть на $[x_{i-1}, x_i]$ расположен $(m+1)$ узел, каждый из них соответствует различным $(m+1)$ значениям параметра t : $t_0, t_1, \dots, t_m \in [-1, 1]$.

На каждом отрезке $[x_{i-1}, x_i]$ занумеруем узлы интерполяции:

$$z_k = x_{i-1/2} + t_k h_i / 2, \quad k = 0, 1, \dots, m,$$

где $x_{i-1/2} = (x_{i-1} + x_i) / 2$ — середина элементарного отрезка. На этом отрезке интерполяционный многочлен степени m в форме Лагранжа:

$$L_m(x) = \sum_{k=0}^m f(z_k) \prod_{j=0, k \neq j}^m \frac{x - z_j}{z_k - z_j}.$$

Вычислим интеграл от L_m на отрезке $[x_{i-1}, x_i]$, используя замену переменной $x = x_{i-1/2} + th_i / 2$:

$$I_i \approx \int_{x_{i-1}}^{x_i} L_m(x)dx = \sum_{k=0}^m f(z_k) \int_{x_{i-1}}^{x_i} \prod_{j=0, k \neq j}^m \frac{x - z_j}{z_k - z_j} dx = h_i \sum_{k=0}^m f(z_k) a_k,$$

$$a_k = \frac{1}{2} \int_{-1}^1 \prod_{j=0, k \neq j}^m \frac{t - t_j}{t_k - t_j} dt.$$

Для интеграла $I = \int_a^b f(x) = \sum_{i=1}^n I_i$ получили *составную квадратурную формулу интерполяционного типа*:

$$I \approx \sum_{i=1}^n h_i \sum_{k=0}^m f(x_{i-1/2} + t_k h_i / 2) a_k. \quad (5.10)$$

Рассмотренные ранее простейшие квадратурные формулы являются частными случаями формулы (5.10) для случая равноотстоящих значений t_0, t_1, \dots, t_m , которые в этом случае относятся к классу *формул Ньютона – Котеса*

Частный случай формулы Ньютона – Котеса при $m = 3$ – *правило 3/8*:

$$I \approx \sum_{i=1}^n \frac{h_i}{8} \left\{ f(x_{i-1}) + 3f\left(x_{i-1} + \frac{h_i}{3}\right) + 3f\left(x_i - \frac{h_i}{3}\right) + f(x_i) \right\};$$

и $m = 4$ – *формула Милана*:

$$I \approx \sum_{i=1}^n \frac{h_i}{90} \left\{ 7f(x_{i-1}) + 32f\left(x_{i-1} + \frac{h_i}{4}\right) + 12f(x_{i-1/2}) + 32f\left(x_i - \frac{h_i}{4}\right) + 7f(x_i) \right\}.$$

Приведем без доказательства теорему об оценке погрешности формулы (5.10).

Теорема. Пусть функция f на (a, b) имеет непрерывную производную порядка $(m + 1)$. Тогда для формулы (5.10) справедлива оценка погрешности:

$$\left| I - I^h \right| \leq A_{m+1} \frac{M_{m+1}(b-a)}{2^{m+2}(m+1)!} h^{m+1},$$

где

$$A_{m+1} = \int_{-1}^1 \left| \prod_{k=0}^m (t - t_k) \right| dt.$$

Замечание 3. Квадратурные формулы интерполяционного типа точны для всех многочленов степени m .

Рассмотрим *обусловленность квадратурных формул интерполяционного типа*. Определим абсолютную погрешность функции $f^*(x)$ равенством $\bar{\Delta}(f^*) = \sup_{x \in [a, b]} |f^*(x) - f(x)|$.

Поскольку

$$\Delta(I^*) = \left| I - I^* \right| = \left| \int_a^b (f(x) - f^*(x)) dx \right| \leq (b-a) \bar{\Delta}(f^*),$$

то абсолютное число обусловленности этой задачи равно $b - a$, и задача вычисления интеграла от приближенно заданной функции является устойчивой.

Рассмотрим чувствительность общей формулы $\int_a^b f(x) \approx \sum_{i=0}^n A_i f(x_i)$ к погрешностям задания функции f .

Очевидно, что

$$\begin{aligned} \Delta(I^*) &= |I - I^*| = \left| \int_a^b (f(x) - f^*(x)) dx \right| \leq \\ &\leq \left| \sum_{i=0}^n A_i f(x_i) - \sum_{i=0}^n A_i f^*(x_i) \right| \leq \sum_{i=0}^n |A_i| \bar{\Delta}(f^*). \end{aligned}$$

Таким образом, абсолютное число обусловленности равно $\nu_{\Delta} = \sum_{i=0}^n |A_i|$. Все квадратурные формулы точны для многочленов нулевой степени и поэтому

$$b - a = \int_a^b 1 dx = \sum_{i=0}^n A_i.$$

Следовательно, если все веса A_i – положительные, то

$$\nu_{\Delta} = \sum_{i=0}^n |A_i| = \sum_{i=0}^n A_i = b - a.$$

При больших m среди весов формулы (5.10) появляются отрицательные, и значения $\nu_{\Delta} = \sum_{i=0}^n |A_i|$ становятся большими. При $m = 10$ число обусловленности $\nu_{\Delta} \approx 3,1(b - a)$, при $m = 30$ число $\nu_{\Delta} \approx 560(b - a)$. В силу плохой обусловленности эти формулы при $m \geq 10$ не используются.

5.4. Квадратурные формулы Гаусса. Полиномы Лежандра

Квадратурная формула $\int_a^b f(x) dx \approx \sum_{i=0}^n A_i f(x_i)$, построенная интегрированием интерполяционного многочлена степени n с фиксированными узлами x_0, x_1, \dots, x_n , точна для всех многочленов степени n . Рассмотрим задачу построения квадратурной формулы, точной для многочленов наиболее высокой степени, чем n , при заданном количестве $(n + 1)$ узлов за счет выбора узлов.

Рассмотрим построение *квадратурной формулы Гаусса* сначала для отрезка $[-1,1]$.

$$\int_{-1}^1 f(t)dt \approx \sum_{i=0}^n A_i f(t_i). \quad (5.11)$$

Переход к произвольному отрезку $[a,b]$ осуществляется с помощью замены переменной $x = (a+b)/2 + t(b-a)/2$.

$$\int_a^b f(x)dx \approx \frac{b-a}{2} \sum_{i=0}^n A_i f((a+b)/2 + t_i(b-a)/2). \quad (5.12)$$

Формула (5.11) точна для многочленов степени m , тогда и только тогда, когда она точна для всех функций: $1, t, t^2, \dots, t^m$. Это означает, что узлы $t_0, t_1, \dots, t_n \in [-1,1]$ и веса A_i должны удовлетворять системе нелинейных уравнений

$$\int_{-1}^1 f(t)dt \approx \sum_{i=0}^n A_i t_i^k = \int_{-1}^1 t^k dt = \frac{1 - (-1)^{k+1}}{k+1}, \quad i=0, 1, \dots, n. \quad (5.13)$$

Система (5.13) имеет единственное решение $t_0, t_1, \dots, t_n \in [-1,1]$, $A_i, 0 \leq i \leq n$, когда $m+1 = 2n+2$ или $m = 2n+1$.

Пример 1. Построим квадратурную формулу Гаусса с двумя узлами $t_0, t_1 \in [-1,1]$. В этом случае $m=3$, $n=1$ и система (5.13) при $k=0,1$ примет вид

$$\begin{aligned} A_0 + A_1 &= \int_{-1}^1 t^0 dt = 2, & A_0 t_0 + A_1 t_1 &= \int_{-1}^1 t dt = 0, \\ A_0 t_0^2 + A_1 t_1^2 &= \int_{-1}^1 t^2 dt = \frac{2}{3}, & A_0 t_0^3 + A_1 t_1^3 &= \int_{-1}^1 t^3 dt = 0. \end{aligned}$$

Решая ее, находим значения $A_0 = A_1 = 1$, $t_0 = -1/\sqrt{3}$, $t_1 = 1/\sqrt{3}$. Таким образом, получили *квадратурную формулу Гаусса с двумя узлами*, точную для многочленов третьей степени:

$$\int_a^b f(x)dx \approx \frac{b-a}{2} \left[f\left(\frac{a+b}{2} - \frac{1}{\sqrt{3}}\frac{b-a}{2}\right) + f\left(\frac{a+b}{2} + \frac{1}{\sqrt{3}}\frac{b-a}{2}\right) \right].$$

Для квадратурной формулы Гаусса справедлива следующая оценка погрешности:

$$\left| \int_a^b f(x) dx - \sum_{i=0}^n A_i f(x_i) \right| \leq \frac{[(n+1)!]^4}{(2n+3)[(2n+2)!]^3} M_{2n+2} (b-a)^{2n+3}.$$

Уже при небольшом числе узлов ($4 \leq n \leq 10$) формула Гаусса обеспечивает очень высокую точность для достаточно гладких функций.

Решить систему (5.13) при произвольном n достаточно сложно, но с помощью полиномов Лежандра, можно вычислить узлы $t_0, t_1, \dots, t_{n-1} \in [-1, 1]$, как корни полиномов. Полиномы Лежандра определяются формулой

$$P_n(t) = \frac{1}{2^n n!} \frac{d^n}{dt^n} (t^2 - 1)^n. \quad (5.14)$$

Первые полиномы имеют вид

$$P_0(t) = 1, \quad P_1(t) = t, \quad P_2(t) = \frac{3}{2}t^2 - \frac{1}{2}, \quad P_3(t) = \frac{5}{2}t^3 - \frac{3}{2}t.$$

Полиномы Лежандра обладают рядом замечательных свойств, из которых следует:

1. Узлы $t_0, t_1, \dots, t_{n-1} \in [-1, 1]$ квадратурной формулы Гаусса с n узлами определяются как корни полинома Лежандра степени n ;
2. Веса A_i из формулы (5.11) находятся из интегральных формул

$$A_i = \int_{-1}^1 \frac{(t-t_0)\dots(t-t_{i-1})(t-t_{i+1})\dots(t-t_{n-1})}{(t_i-t_0)\dots(t_i-t_{i-1})(t_i-t_{i+1})\dots(t_i-t_{n-1})} dt. \quad (5.15)$$

Пример 2. Построим квадратурную формулу Гаусса с тремя узлами. В этом случае $m = 5$, $n = 2$.

Узлы $t_0 = -\sqrt{3/5}$, $t_1 = 0$, $t_2 = \sqrt{3/5}$ находим как корни полинома третьей степени $P_3(t) = \frac{5}{2}t^3 - \frac{3}{2}t = 0$.

Остается вычислить весовые коэффициенты:

$$A_0 = \int_{-1}^1 \frac{t(t - \sqrt{3/5})}{(-\sqrt{3/5})(-2\sqrt{3/5})} dt = 5/9,$$

$$A_1 = \int_{-1}^1 \frac{(t + \sqrt{3/5})(t - \sqrt{3/5})}{(-3/5)} dt = 8/9,$$

$$A_2 = \int_{-1}^1 \frac{t(t - \sqrt{3/5})}{\sqrt{3/5}(2\sqrt{3/5})} dt = 5/9.$$

В результате квадратурная формула Гаусса на интервале $[-1,1]$ с тремя узлами запишется в виде

$$\int_{-1}^1 f(t) dt \approx \sum_{i=0}^2 A_i f(t_i) = 5/9 f(-\sqrt{3/5}) + 8/9 f(0) + 5/9 f(\sqrt{3/5}).$$

Она является точной для многочлена пятой степени.

Квадратурная формула Гаусса с тремя узлами на произвольном отрезке $[a,b]$ осуществляется с помощью замены переменной $x = (a+b)/2 + t(b-a)/2$ и имеет вид

$$\int_a^b f(x) dx \approx \frac{b-a}{2} [5/9 f((a+b)/2 - \sqrt{3/5}) + 8/9 f((a+b)/2) + 5/9 f((a+b)/2 + \sqrt{3/5})].$$

Квадратурная формула Гаусса с четырьмя узлами на произвольном отрезке $[a,b]$ является точной для многочлена седьмой степени. Укажем узлы и соответствующие веса этой формулы:

$$\begin{aligned} t_0 &= -0,8611363115, & A_0 &= 0,3478548451; \\ t_1 &= -0,3399810436, & A_1 &= 0,6521451549; \\ t_2 &= 0,3399810436, & A_2 &= 0,6521451549; \\ t_4 &= 0,8611363115, & A_4 &= 0,3478548451. \end{aligned}$$

Замечание 4. Весовые коэффициенты квадратурной формулы Гаусса всегда положительны, что гарантирует хорошую обусловленность формулы с абсолютным числом

$$\nu_{\Delta} = \sum_{i=0}^n |A_i| = \sum_{i=0}^n A_i = b - a.$$

5.5. Апостериорные оценки погрешности

Применение неравенств вида (5.7)–(5.9) для априорной оценки погрешности в большинстве случаев неэффективно или невозможно. Это связано с трудностью оценки производных подынтегральной функции и с тем, что полученные оценки, как правило, бывают сильно завышенными.

Пусть I^h – приближенное значение интеграла $\int_a^b f(x)dx$, вычисленное по некоторой квадратурной формуле (5.10) с постоянным шагом $h = (b - a)/n$, либо с $h = \max h_i$ для случая переменного шага. Предположим, что для погрешности этой формулы справедливо представление

$$I - I^h = Ch^k + o(h^k), \quad (5.16)$$

где $C \neq 0$, и $k > 0$.

Тогда величина Ch^k называется *главным членом погрешности* квадратурной формулы, и число k представляет порядок точности формулы. При достаточно малом h для погрешности в предположении (5.16) справедливо приближенное равенство

$$I - I^h \approx Ch^k. \quad (5.17)$$

Уменьшение шага в два раза приводит к уменьшению погрешности в 2^k раз:

$$I - I^{h/2} \approx \frac{1}{2^k} Ch^k \approx \frac{1}{2^k} (I - I^h). \quad (5.18)$$

Вычитая из равенства (5.17) равенство (5.18), получаем

$$I^{h/2} - I^h \approx Ch^k - \frac{1}{2^k} Ch^k = \frac{1}{2^k} Ch^k (2^k - 1).$$

Учитывая формулу (5.18), имеем

$$I^{h/2} - I^h \approx (I - I^{h/2})(2^k - 1).$$

Откуда

$$I - I^{h/2} \approx \frac{I^{h/2} - I^h}{2^k - 1}. \quad (5.19)$$

Формула (5.19) дает *апостериорную* (на основании опыта) оценку погрешности значения $I^{h/2}$:

$$\left| \frac{I^{h/2} - I^h}{2^k - 1} \right| < \varepsilon,$$

которую принято называть *правилом двойного пересчета* или *правилом Рунге*.

Для формул прямоугольника и трапеций (5.7)–(5.8) $k = 2$, поэтому формула (5.19) для апостериорной оценки погрешности принимает следующий вид:

$$\left| I - I_{\text{пр}}^{h/2} \right| \approx \left| \frac{I_{\text{пр}}^{h/2} - I_{\text{пр}}^h}{3} \right| < \varepsilon, \quad \left| I - I_{\text{тр}}^{h/2} \right| \approx \left| \frac{I_{\text{тр}}^{h/2} - I_{\text{тр}}^h}{3} \right| < \varepsilon.$$

Для формулы Симпсона (5.9) $k = 4$, поэтому по формуле (5.19) имеем

$$\left| I - I_C^{h/2} \right| \approx \left| \frac{I_C^{h/2} - I_C^h}{15} \right| < \varepsilon.$$

Для формулы Милана $k = 5$, поэтому по формуле (5.19) имеем

$$\left| I - I_M^{h/2} \right| \approx \left| \frac{I_M^{h/2} - I_M^h}{31} \right| < \varepsilon.$$

5.6. Построение первообразной с помощью численного интегрирования

Рассмотрим задачу по известной функции $f(x)$ на отрезке $[a, b]$, восстановить ее первообразную

$$y(x) = \int_a^x f(\xi) d\xi, \quad x \in [a, b] \quad (5.20)$$

При каждом фиксированном x функцию $y(x)$ можно рассматривать как определенный интеграл и вычислять с помощью одного из известных методов. Однако, если необходимо определить значения функции $y(x)$ в большом количестве точек, то целесообразнее разбить отрезок $[a, b]$ на элементарные отрезки точками $a = x_0 < x_1 < \dots < x_n = b$, а затем составить таблицу значений $y_i \approx y(x_i)$, $0 \leq i \leq n$ по формуле

$$y_i = y_{i-1} + I_i^{h_i}, \quad 1 \leq i \leq n.$$

Здесь $y_0 = 0$, а $I_i^{h_i}$ – приближенное значение интеграла $\int_{x_{i-1}}^{x_i} f(\xi) d\xi$, полученное с помощью одной из квадратурных формул.

Функцию $y(x)$, можно в любой из промежуточных точек восстановить, используя интерполяцию. Так как значения производной функции $y(x)$ известно, $y'(x_i) = f(x_i)$, то для интерполяции на каждом элементарном отрезке $[x_{i-1}, x_i]$ хорошо подходит кубический многочлен Эрмита.

Замечание 5. Существует большой класс функций, которые не могут быть представлены через элементарные функции:

$$Si(x) = \int_0^x \frac{\sin(t)}{t} dt, \quad erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \text{ и др.}$$

Методы численного интегрирования позволяют вычислять значения такого рода функций при любом x . Графики функций $Si(x)$ и $erf(x)$ представлены на рис. 22 и 23.

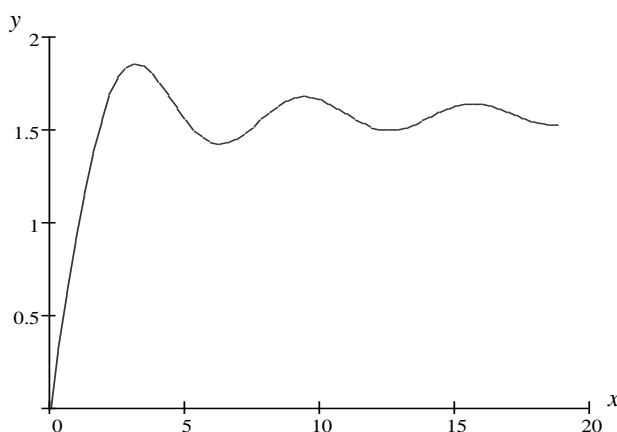


Рис. 22

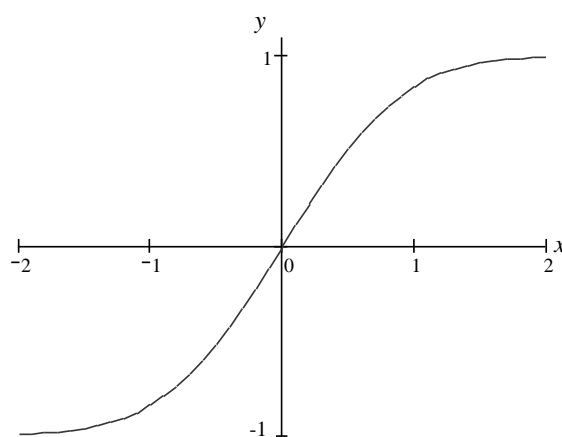


Рис. 23

6. ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ ЗАДАЧИ КОШИ

Наиболее универсальные методы решения задачи Коши — это разностные методы, когда производные заменяются разностными отношениями. В результате исходное дифференциальное уравнение сводится к системе алгебраических уравнений, которые называются разностными. Решение этой системы дает приближенное решение исходной задачи.

Рассмотрим задачу Коши для дифференциального уравнения первого порядка:

$$y' = f(x, y(x)), \quad y(x_0) = y_0. \quad (6.1)$$

Если функция $f(x, y)$ непрерывна и удовлетворяет условию Липшица по переменной y в некоторой окрестности точки (x_0, y_0) , то можно указать такой отрезок $[a, b]$, $x_0 \in [a, b]$, на котором решение задачи существует и является единственным.

6.1. Разностная аппроксимация производных

Пусть на отрезке $[a, b]$ задан набор точек

$$a = x_0 < x_1 < \dots < x_i < x_{i+1} < \dots < x_n = b.$$

Будем называть его сеткой. Пусть сетка равномерная, т. е.

$$x_{i+1} - x_i = h = (b - a)/n, \quad 0 \leq i \leq n - 1.$$

Пусть каждой точке сетки x_i сопоставлено по определенному закону число y_i . Совокупность этих чисел назовем сеточной функцией $\mathbf{y} = \{y_0, y_1, y_2, \dots, y_n\}$. Сеточные функции, определенные на сетке, образуют $(n + 1)$ -мерное линейное пространство. Чтобы сравнить сеточные функции между собой введем в рассмотрение норму:

$$\|\mathbf{y}\|_C = \max_{0 \leq i \leq n} |y_i|. \quad (6.2)$$

Вместо производных первого порядка для сеточных функций вводятся разностные отношения:

$$L_h^+(y_i) = \frac{y_{i+1} - y_i}{h}, \quad 0 \leq i \leq n-1; \quad (6.3)$$

$$L_h^-(y_i) = \frac{y_i - y_{i-1}}{h}, \quad 1 \leq i \leq n; \quad (6.4)$$

$$L_h^0(y_i) = \frac{y_{i+1} - y_{i-1}}{2h}, \quad 1 \leq i \leq n-1. \quad (6.5)$$

Отношение (6.3) называют правой разностной производной, отношение (6.4) – левой разностной производной, (6.5) – центральной разностной производной.

Вычислим первую производную функции $y'(x_i)$ и сопоставим для каждого разностного отношения (6.3)–(6.4) погрешность аппроксимации производной в точках x_i :

$$\xi_i^+ = L_h^+(y_i) - y'(x_i), \quad 0 \leq i \leq n-1; \quad (6.6)$$

$$\xi_i^- = L_h^-(y_i) - y'(x_i), \quad 1 \leq i \leq n; \quad (6.7)$$

$$\xi_i^0 = L_h^0(y_i) - y'(x_i), \quad 1 \leq i \leq n-1. \quad (6.8)$$

Предположим, что функция $y(x)$ дважды непрерывно дифференцируема на отрезке $[a, b]$, и запишем для нее формулу Тейлора с остаточным членом в форме Лагранжа:

$$y_{i+1} = y(x_i + h) = y_i + y'(x_i)h + \frac{1}{2}y''(x_i + \theta_i h)h^2, \quad \theta_i \in [0, 1]. \quad (6.9)$$

Тогда из формул (6.3), (6.6) находим $\xi_i^+ = \frac{1}{2}y''(x_i + \theta_i h)h$.

Аналогично получаем $\xi_i^- = -\frac{1}{2}y''(x_i - \theta_i h)h$.

Для оценки ξ_i^0 предположим, что функция $y(x)$ трижды непрерывно дифференцируема на отрезке $[a, b]$, и запишем для нее формулу Тейлора с остаточным членом в форме Лагранжа, добавляя еще один член:

$$y_{i+1} = y(x_i + h) = y_i + y'(x_i)h + \frac{1}{2}y''(x_i)h^2 + \frac{1}{6}y'''(x_i + \theta_{1i}h)h^3,$$

$$y_{i-1} = y(x_i - h) = y_i - y'(x_i)h + \frac{1}{2}y''(x_i)h^2 - \frac{1}{6}y'''(x_i - \theta_{2i}h)h^3.$$

Из формул (6.5), (6.8) будем иметь

$$\xi_i^0 = \frac{1}{12}(y'''(x_i + \theta_{1i}h) - y'''(x_i - \theta_{2i}h))h^2.$$

Пусть для функции $f(x)$, k раз непрерывно дифференцируемой, определена величина $M_k = \max_{[a,b]} |f^{(k)}(x)|$, тогда для оценок погрешности аппроксимации первых производных в точках x_i будем иметь

$$|\xi_i^+| \leq \frac{1}{2}M_2h, \quad |\xi_i^-| \leq \frac{1}{2}M_2h, \quad |\xi_i^0| \leq \frac{1}{6}M_3h^2. \quad (6.10)$$

Таким образом, центральная разностная производная дает более хороший результат: она аппроксимирует производную $y(x)$ в точках x_i со вторым порядком точности относительно h для функций, трижды непрерывно дифференцируема на отрезке $[a,b]$.

Для разностной аппроксимации второй производной составим разностное отношение первых разностных производных:

$$L_h(y_i) = \frac{(y_{i+1} - y_i)/h - (y_i - y_{i-1})/h}{h} = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2}, \quad 1 \leq i \leq n-1. \quad (6.11)$$

Для оценки погрешности аппроксимации второй производной в точках x_i будем иметь

$$|\xi_i| \leq \frac{1}{12}M_4h^2. \quad (6.12)$$

6.2. Метод Эйлера

Пусть нам нужно построить решение задачи (6.1) на отрезке $[a,b]$, $x_0 = a$. Введем шаг $h = (b-a)/n$ и образуем сетку:

$$x_i = x_0 + ih, \quad 0 \leq i \leq n.$$

Сопоставим задаче (6.1) на отрезке разностную задачу

$$(u_{i+1} - u_i)/h = f(x_i, u_i), \quad 0 \leq i \leq n-1; \quad u_0 = y_0. \quad (6.13)$$

Уравнение является разностным уравнением первого порядка, которое называется *схемой Эйлера*. Его можно переписать в виде *явной* рекуррентной схемы:

$$u_{i+1} = u_i + f(x_i, u_i)h \quad 0 \leq i \leq n-1; \quad (6.14)$$

Для оценки точности решения исходной задачи Коши (6.1) функции $y(x)$ образуем две сеточные функции – \mathbf{z} и ψ :

$$z_i = u_i - y_i, \quad 0 \leq i \leq n;$$

$$\psi_i = \frac{y_{i+1} - y_i}{h} - f(x_i, y_i), \quad 0 \leq i \leq n-1.$$

Сеточную функцию \mathbf{z} называют *погрешностью решения*. Сеточную функцию ψ называют *погрешностью аппроксимации* схемы на решении $y(x)$. Погрешность аппроксимации схемы была бы равна нулю, если бы уравнения (6.1) и (6.13) совпадали.

Подставим выражение $u_i = z_i + y_i$ в уравнение (6.13), в результате получим

$$(y_{i+1} - y_i)/h + (z_{i+1} - z_i)/h = f(x_i, y_i + z_i),$$

или

$$(z_{i+1} - z_i)/h = f(x_i, y_i + z_i) - f(x_i, y_i) - \left\{ (y_{i+1} - y_i)/h - f(x_i, y_i) \right\} =$$

$$= f(x_i, y_i + z_i) - f(x_i, y_i) - \psi_i.$$

Разность значений функции $f(x, y)$ с помощью формулы Лагранжа можно представить как $f(x_i, y_i + z_i) - f(x_i, y_i) = \frac{\partial f}{\partial y}(x_i, y_i + \theta_i z_i)z_i$, и записать для z_i рекуррентную формулу с нулевым $z_0 = u_0 - y_0 = 0$ начальным значением:

$$z_{i+1} = \left(1 + h \frac{\partial f}{\partial y}(x_i, y_i + \theta_i z_i)\right) z_i - \psi_i h. \quad (6.15)$$

Пусть в рассматриваемой области решения $\left| \frac{\partial f}{\partial y}(x, y) \right| \leq M$. Это позволяет записать довольно грубую оценку

$$\left| 1 + h \frac{\partial f}{\partial y}(x_i, y_i + \theta_i z_i) \right| \leq 1 + Mh < e^{Mh} = q, \quad q > 1.$$

Тогда из (6.15) следуют рекуррентные неравенства

$$|z_{i+1}| \leq q|z_i| + \|\Psi\|_C h,$$

которые дают цепочку оценок:

$$\begin{aligned} z_0 &= 0, \\ |z_1| &\leq \|\Psi\|_C h, \\ |z_2| &\leq (1+q)\|\Psi\|_C h, \\ |z_3| &\leq (1+q+q^2)\|\Psi\|_C h, \\ &\dots\dots\dots \\ |z_n| &\leq (1+q+q^2+\dots+q^{n-1})\|\Psi\|_C h, \end{aligned}$$

где $1+q+q^2+\dots+q^{n-1} < nq^n = ne^{Mhn} = ne^{M(b-a)}$.

Для каждой погрешности z_i получили оценку

$$|z_i| \leq nhe^{M(b-a)}\|\Psi\|_C = (b-a)e^{M(b-a)}\|\Psi\|_C,$$

с учетом рассматриваемой нормы оценка погрешности решения примет вид

$$\|\mathbf{z}\|_C \leq (b-a)e^{M(b-a)}\|\Psi\|_C. \quad (6.16)$$

Из оценки (6.16) следует, что чем лучше разностное уравнение аппроксимирует дифференциальное, тем меньше погрешность решения.

Для погрешности аппроксимации в силу первого неравенства (6.10) имеем $\|\Psi\|_C \leq \frac{1}{2}M_2h$, что позволяет переписать оценку (6.16) в форме

$$\|\mathbf{z}\|_C \leq \frac{1}{2}(b-a)e^{M(b-a)}M_2h. \quad (6.17)$$

Неравенство (6.17) показывает, что погрешность решения стремится к нулю со скоростью h . В связи с этим метод Эйлера называют *методом первого порядка точности* относительно шага h . Геометрическая иллюстрация явного метода Эйлера представлена на рис. 24.

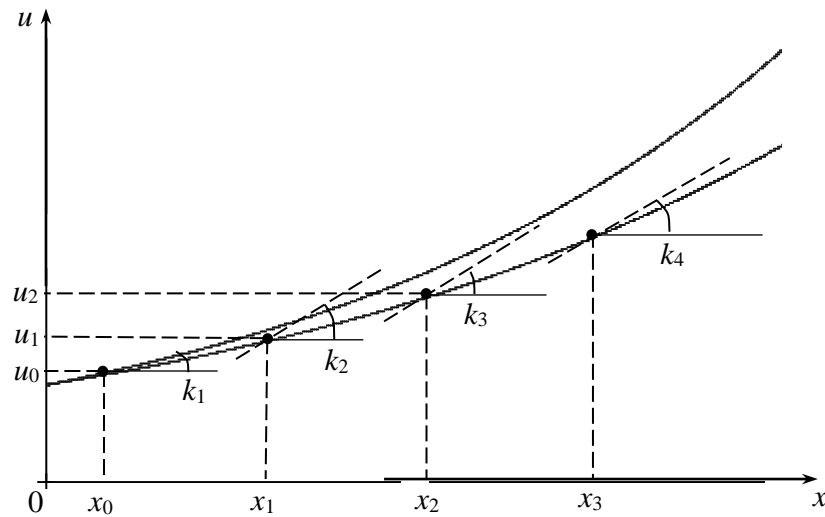


Рис. 24

6.3. Метод Рунге – Кутты

Из оценки (6.17) метода Эйлера можно сделать вывод, чтобы повысить точность метода, нужно улучшить аппроксимацию дифференциального уравнения разностной схемой. Запишем для решения уравнения (6.1) формулу Тейлора с остаточным членом в форме Лагранжа:

$$y_{i+1} = y(x_i + h) = y_i + y'(x_i)h + \frac{1}{2} y''(x_i)h^2 + \frac{1}{6} y'''(x_i)h^3 + \dots \quad (6.18)$$

Оставим в этом разложении слагаемые до порядка h^2 включительно и заменим $y'(x_i)$ и $y''(x_i)$ в силу уравнения (6.1) формулами

$$y'(x_i) = f(x_i, y_i), \quad y''(x_i) = \frac{\partial f}{\partial x}(x_i, y_i) + \frac{\partial f}{\partial y}(x_i, y_i) f(x_i, y_i).$$

В результате получим новое приближенное рекуррентное соотношение, более сложное, чем (6.14):

$$u_{i+1} = u_i + f(x_i, u_i)h + \frac{1}{2} \left(\frac{\partial f}{\partial x}(x_i, u_i) + \frac{\partial f}{\partial u}(x_i, u_i) f(x_i, u_i) \right) h^2,$$

которое можно записать в виде разностного уравнения:

$$\frac{u_{i+1} - u_i}{h} = f(x_i, u_i) + \frac{1}{2} \left(\frac{\partial f}{\partial x}(x_i, u_i) + \frac{\partial f}{\partial u}(x_i, u_i) f(x_i, u_i) \right) h. \quad (6.19)$$

При разработке разностных схем высокого порядка точности стремятся заменить вычисление производных функции нахождением значений самой функции в нескольких точках.

Замечание. Если в разложении (6.18) оставить члены порядка h^3 или h^4 , то получатся разностные схемы третьего или четвертого порядка точности.

Основная идея метода Рунге – Кутты состоит в том, чтобы заменить правую часть уравнения (6.19) на сумму значений функции f и двух различных точках с точностью до членов порядка h^2 . С этой целью положим

$$\begin{aligned} f(x_i, u_i) + \frac{1}{2} \left(\frac{\partial f}{\partial x}(x_i, u_i) + \frac{\partial f}{\partial u}(x_i, u_i) f(x_i, u_i) \right) h = \\ = \beta f(x_i, u_i) + \alpha f(x_i + \gamma h, u_i + \delta h) + O(h^2), \end{aligned}$$

где $\alpha, \beta, \gamma, \delta$ – четыре свободных параметра, подлежащих определению.

Разложим функцию $f(x_i + \gamma h, u_i + \delta h)$ по степеням h :

$$f(x_i + \gamma h, u_i + \delta h) = f(x_i, u_i) + \left(\gamma \frac{\partial f}{\partial x}(x_i, u_i) + \delta \frac{\partial f}{\partial u}(x_i, u_i) \right) h + O(h^2).$$

Тогда

$$\begin{aligned} f(x_i, u_i) + \frac{1}{2} \left(\frac{\partial f}{\partial x}(x_i, u_i) + \frac{\partial f}{\partial u}(x_i, u_i) f(x_i, u_i) \right) h = \beta f(x_i, u_i) + \\ + \alpha \left(f(x_i, u_i) + \left(\gamma \frac{\partial f}{\partial x}(x_i, u_i) + \delta \frac{\partial f}{\partial u}(x_i, u_i) \right) h + O(h^2) \right) + O(h^2), \end{aligned}$$

приравнявая слева и справа члены, содержащие h , и члены, не содержащие h , получим три уравнения:

$$\alpha + \beta = 1, \quad \alpha\gamma = 1/2, \quad \alpha\delta = f(x_i, u_i)/2.$$

Они позволяют выразить параметры β, γ, δ через α :

$$\beta = 1 - \alpha, \quad \gamma = 1/(2\alpha), \quad \delta = f(x_i, u_i)/(2\alpha).$$

Уравнение (6.19) можно записать в форме рекуррентного однопараметрического семейства:

$$u_{i+1} = u_i + \left[(1 - \alpha) f(x_i, u_i) + \alpha f\left(x_i + \frac{h}{2\alpha}, u_i + \frac{h}{2\alpha} f(x_i, u_i)\right) \right] h. \quad (6.20)$$

Наиболее распространенные разностные схемы этого семейства *схем Рунге – Кутты* соответствуют двум значениям параметра α : $\alpha = 1/2$ и $\alpha = 1$. При $\alpha = 1/2$ формула (6.20) принимает вид

$$u_{i+1} = u_i + \frac{h}{2} [f(x_i, u_i) + f(x_i + h, u_i + hf(x_i, u_i))]. \quad (6.21)$$

Ее можно разбить на два действия «предиктор – корректор»:

$$\tilde{u}_{i+1} = u_i + hf(x_i, u_i), \quad u_{i+1} = u_i + \frac{h}{2} [f(x_i, u_i) + f(x_{i+1}, \tilde{u}_{i+1})]. \quad (6.22)$$

Первое действие происходит по схеме Эйлера – это грубое предсказание результата, второе действие является уточнением результата, его коррекцией.

При $\alpha = 1$ формула (6.20) принимает вид

$$u_{i+1} = u_i + hf\left(x_i + \frac{h}{2}, u_i + \frac{h}{2} f(x_i, u_i)\right). \quad (6.23)$$

Эту формулу также можно разбить на два действия «предиктор – корректор»:

$$\tilde{u}_{i+1/2} = u_i + \frac{h}{2} f(x_i, u_i), \quad u_{i+1} = u_i + hf\left(x_i + \frac{h}{2}, \tilde{u}_{i+1/2}\right).$$

Как и в случае метода Эйлера, можно получить оценку погрешности решения вида

$$\|\mathbf{z}\|_C \leq \frac{1}{2} (b - a) e^{M(b-a)} Kh^2. \quad (6.24)$$

Неравенство (6.24) показывает, что погрешность решения стремится к нулю со скоростью h^2 . В связи с этим метод Рунге – Кутты называют *методом второго порядка точности* относительно шага h .

Наиболее часто при проведении реальных расчетов используется *схема Рунге – Кутты четвертого порядка точности*:

$$\frac{u_{i+1} - u_i}{h} = \frac{1}{6} (k_1 + 2k_2 + 2k_3 + k_4), \quad (6.25)$$

где

$$k_1 = f(x_i, u_i), \quad k_2 = f\left(x_i + \frac{h}{2}, u_i + \frac{h}{2} k_1\right), \\ k_3 = f\left(x_i + \frac{h}{2}, u_i + \frac{h}{2} k_2\right), \quad k_4 = f\left(x_i + h, u_i + hk_3\right).$$

Предположим, что для погрешности явного одношагового метода справедливо следующее представление для оценки погрешности:

$$\|\mathbf{z}\|_C = Ch^k + o(h^k),$$

где $C \neq 0$, и $k > 0$.

Тогда метод имеет порядок точности, равный k , а локальная погрешность $z_i = u_i - y_i$ допускает представление

$$|z_i|_C = Lh^{k+1} + o(h^{k+1}).$$

Уменьшим шаг интегрирования вдвое и вычислим приближение к значению решения в точке x_{i+1} с помощью того же одношагового метода, но уже с помощью двух элементарных действий с половинным шагом. Например, по формуле (6.21) имеем

$$u_{i+1/2}^{h+1/2} = u_i + \frac{h}{4} [f(x_i, u_i) + f(x_i + h/2, u_i + h/2 f(x_i, u_i))],$$

$$u_{i+1}^{h+1/2} = u_{i+1/2}^{h+1/2} + \frac{h}{4} [f(x_{i+1/2}, u_{i+1/2}^{h+1/2}) + f(x_{i+1/2} + h/2, u_{i+1/2}^{h+1/2} + h/2 f(x_{i+1/2}, u_{i+1/2}^{h+1/2}))].$$

Тогда, аналогично формуле двойного пересчета (7.19), можно получить апостериорную оценку локальной погрешности значения $u_{i+1}^{h+1/2}$:

$$y_{i+1} - u_{i+1}^{h+1/2} \approx \frac{u_{i+1}^{h+1/2} - u_{i+1}^h}{2^k - 1} = \xi_{i+1}. \quad (6.26)$$

Из последнего приближенного равенства имеем апостериорную оценку локальной погрешности значения $u_{i+1}^{h+1/2}$.

Уточненное значение $u_{i+1} = u_{i+1}^{h+1/2} + \xi_{i+1}$ в дальнейшем при расчетах принимается за приближенное решение задачи Коши в точке x_{i+1} .

Заметим, что этот способ контроля точности приводит к увеличению времени счета на 50 %.

6.4. Метод Адамса

Из уравнения (6.1) находим

$$y_{i+1} = y_i + \int_{x_i}^{x_{i+1}} f(x, y(x)) dx. \quad (6.27)$$

Это соотношение непосредственно использовать нельзя, т. к. при переходе от точки сетки x_i к точке x_{i+1} подынтегральная функция неизвестна.

Рассмотрим, как эта проблема решается в методе Адамса. Пусть в процессе численного решения задачи расчет доведен до точки x_i и известны величины y_i и $f(x_i, y_i)$. Построим интерполяционный многочлен $P_m(x)$ некоторой фиксированной степени m , принимающий в точках x_j , $i-m \leq j \leq i$ значения $f(x_j, y_j)$.

Полином можно записать в форме Лагранжа

$$P_m(x) = \sum_{j=i-m}^m f(x_j, y_j) \frac{(x-x_{i-m})(x-x_{i-m+1})\dots(x-x_{j-1})(x-x_{j+1})\dots(x-x_m)}{(x_j-x_{i-m})(x_j-x_{i-m+1})\dots(x_j-x_{j-1})(x_j-x_{j+1})\dots(x_j-x_m)}.$$

Многочлен $P_m(x)$ легко интегрируется, поэтому формула (6.27) приводит к рекуррентной формуле

$$u_{i+1} = u_i + \int_{x_i}^{x_{i+1}} P_m(x) dx = u_i + \sum_{j=i-m}^i a_j f(x_j, u_j), \quad (6.28)$$

где коэффициенты

$$a_j = \int_{x_i}^{x_{i+1}} \frac{(x-x_{i-m})(x-x_{i-m+1})\dots(x-x_{j-1})(x-x_{j+1})\dots(x-x_m)}{(x_j-x_{i-m})(x_j-x_{i-m+1})\dots(x_j-x_{j-1})(x_j-x_{j+1})\dots(x_j-x_m)} dx.$$

Метод Адамса (6.28) является линейным *многошаговым* методом. При применении формулы (6.28) необходимо использовать $(m+1)$ предыдущее значение дискретной функции \mathbf{u} .

Рассмотрим случай полинома первой степени:

$$P_1(x) = f(x_i, y_i) \frac{x-x_{i-1}}{h} - f(x_{i-1}, y_{i-1}) \frac{x-x_i}{h}.$$

Подставляя в интегральную формулу (6.27) и интегрируя, получаем двухшаговую явную схему Адамса второго порядка:

$$u_{i+1} = u_i + \frac{h}{2} [3f(x_i, u_i) - f(x_{i-1}, u_{i-1})] \quad (6.29)$$

Вычислить u_1 по формуле (6.29) нельзя, это значение приходится вычислять каким-то другим методом, напр., методом Рунге – Кутты.

Замечание. Разностное уравнение метода Адамса, соответствующее случаю полинома степени m аппроксимирует уравнение с точностью порядка h^{m+1} относительно h .

Если применить интерполяционный многочлен третьей степени на равномерной сетке из четырех узлов $x_i, x_{i-1}, x_{i-2}, x_{i-3}$ и провести интегрирование (6.28), то рекуррентная явная формула метода Адамса четвертого порядка примет вид

$$u_{i+1} = u_i + \frac{h}{24} [55f(x_i, u_i) - 59f(x_{i-1}, u_{i-1}) + 37f(x_{i-2}, u_{i-2}) - 9f(x_{i-3}, u_{i-3})] \quad (6.30)$$

Вычислить первые три значения u_1, u_2, u_3 по формуле (6.30) нельзя, эти значения приходится вычислять каким-то другим методом четвертого порядка точности относительно h .

Рассмотрим двухшаговую *неявную схему Адамса*, используя многочлен второго порядка, принимающий в точках x_{i+1}, x_i, x_{i-1} , значения f_{i+1}, f_i, f_{i-1} . Его можно записать в форме Лагранжа

$$P_2(x) = f_{i-1} \frac{(x-x_i)(x-x_{i+1})}{2h^2} + f_i \frac{(x-x_{i-1})(x-x_{i+1})}{(-h^2)} + f_{i+1} \frac{(x-x_{i-1})(x-x_i)}{2h^2}.$$

Из интегральной формулы (6.28) после интегрирования получаем искомую неявную схему:

$$u_{i+1} = u_i + \int_{x_i}^{x_{i+1}} f_{i-1} \frac{(x-x_i)(x-x_{i+1})}{2h^2} + f_i \frac{(x-x_{i-1})(x-x_{i+1})}{(-h^2)} + f_{i+1} \frac{(x-x_{i-1})(x-x_i)}{2h^2} dx, \quad (6.31)$$

$$u_{i+1} = u_i + h(5f(x_{i+1}, u_{i+1}) + 8f(x_i, u_i) - f(x_{i-1}, u_{i-1})).$$

Сложность использования неявных методов Адамса заключается в необходимости решать нелинейное уравнение (6.31) относительно u_{i+1} . Значение u_{i+1} можно найти, используя, напр., метод простой итерации.

Методы Адамса требуют меньшего числа вычислений правой части дифференциального уравнения по сравнению с методами Рунге – Кутты того же порядка точности. Для них существуют эффективные методы апостериорной оценки локальной погрешности. Существенным недостатком метода Адамса является нестандартное начало вычислений.

6.5. Устойчивость численных методов решения задачи Коши

Напомним основные теоремы из курса дифференциальных уравнений, связанных с решением задачи Коши

$$y' = f(t, y(t)), \quad y(t_0) = y_0. \quad (6.32)$$

Теорема. Пусть функция $f(t, y)$ определена и непрерывна в полосе

$$\Pi_T = \{(t, y) : t_0 \leq t \leq T; -\infty < y < \infty\}$$

и удовлетворяет условию Липшица по переменной y для всех $t_0 \leq t \leq T$:

$$\|f(t, y_1) - f(t, y_2)\| \leq L \|y_1 - y_2\|.$$

Тогда для каждого начального значения y_0 существует единственное решение $y(t)$ задачи Коши, определенное при $t \in [t_0, T]$.

Теорема. Пусть выполнены условия теоремы. Далее, пусть $y(t)$ – решение задачи (6.32), а $y^*(t)$ – решение задачи

$$(y^*)' = f(t, y^*(t)) + \beta(t), \quad y^*(t_0) = y_0^*.$$

Тогда справедлива оценка

$$\max_{t_0 \leq t \leq T} |y(t) - y^*(t)| \leq K(T) \left(|y_0 - y_0^*| + \int_{t_0}^T |\beta(t)| dt \right),$$

выражающая устойчивость на конечном отрезке $[t_0, T]$ решение задачи Коши по начальным значениям и правой части. Здесь $K(T) = e^{L(T-t_0)}$.

Величина $K(T) = e^{L(T-t_0)}$ может неограниченно расти с ростом T . Как следствие, при достаточно больших временных отрезках $[t_0, T]$ такая задача является плохо обусловленной.

Рассмотрим задачу (6.31) и задачу

$$(y^*)' = f(t, y^*(t)), \quad y^*(t_0) = y_0^*,$$

где $\varepsilon_0 = y_0 - y_0^*$ есть погрешность начальных значений. Погрешность $\varepsilon(t) = y(t) - y^*(t)$ удовлетворяет уравнению

$$\varepsilon'(t) = y'(t) - (y^*)'(t) = f(t, y(t)) - f(t, y^*(t)).$$

По формуле конечных приращений Лагранжа

$$f(t, y(t)) - f(t, y^*(t)) = f'_y(t, \tilde{y}(t))(y(t) - y^*(t)),$$

где $\tilde{y}(t)$ – некоторое промежуточное значение между $y(t)$ и $y^*(t)$.

Обозначим $\lambda(t) = f'_y(t, \tilde{y}(t))$, тогда

$$\varepsilon'(t) = \lambda(t)\varepsilon(t) \quad \text{и} \quad \varepsilon(t_0) = \varepsilon_0. \quad (6.33)$$

Погрешность решения по начальным значениям определяется формулой

$$\varepsilon(t) = \varepsilon_0 \exp \left\{ \int_{t_0}^t \lambda(\tau) d\tau \right\}.$$

Величина $C(t) = \exp \left\{ \int_{t_0}^t \lambda(\tau) d\tau \right\}$ при $f'_y(t, y(t)) < 0$ с ростом t моно-

тонно убывает и погрешность, внесенная в начальное значение, имеет тенденцию к затуханию.

Пример 1. Рассмотрим задачу Коши $y'(t) = y(t) - \sin t + \cos t$, $y(0) = 0$. Ее решением, как нетрудно проверить, является функция $y(t) = \sin t$.

Внесем в начальное условие погрешность $y(0) = \xi$. Решение задачи Коши при этом начальном условии есть $y(t) = \sin t + \xi e^t$. Погрешность решения при росте t неограниченно возрастает.

Заметим, что $\lambda = f'_y(t, y(t)) = 1 > 0$ и величина $C(t) = \exp \left\{ \int_{t_0}^t \lambda(\tau) d\tau \right\}$

для оценки погрешности решения с ростом t монотонно возрастает.

Для изучения устойчивости по начальным данным задачи (6.1) будем рассматривать *модельное уравнение*

$$y(t) = \lambda y(t), \quad y(0) = y_0;$$

которое моделирует локальное распространение погрешности $\varepsilon(t) = y(t) - y^*(t)$.

Его решение $y(t) = y_0 \exp(\lambda t)$ убывает при $\lambda < 0$, что соответствует условию $f'_y(t, y(t)) < 0$.

Кроме того,

$$|y(t)| \leq |y_0|, \quad \text{при } \lambda \leq 0, \quad \text{для всех } t \geq 0, \quad (6.34)$$

т. е. модельное уравнение устойчиво при $\lambda \leq 0$.

Вводится естественное требование: для разностных схем, аппроксимирующих модельное уравнение, должен выполняться аналог неравенства (6.34)

$$|u_i| \leq |u_0|, \quad i = 1, 2, \dots \quad (6.35)$$

Рассмотрим ряд примеров.

1. Явная схема Эйлера

$$(u_{i+1} - u_i) / \tau = \lambda u_i, \quad u_{i+1} = (1 + \tau\lambda)u_i.$$

Отсюда видно, что условие

$$|u_i| \leq |u_{i-1}| \leq \dots \leq |u_0|, \quad \text{при } \lambda < 0 \quad (6.36)$$

выполнено при $|1 + \tau\lambda| \leq 1$ или $-1 \leq 1 + \tau\lambda \leq 1$. Из последнего неравенства находим, что для устойчивости шаг τ по времени t должен удовлетворять условию $\tau \leq 2/|\lambda|$. Явная схема Эйлера *условно устойчива*.

2. Неявная схема Эйлера

$$(u_{i+1} - u_i) / \tau = \lambda u_{i+1}, \quad u_{i+1} = u_i / (1 - \tau\lambda).$$

Условие (6.36) выполнено при любом τ , т. к. $1/(1 - \tau\lambda) \leq 1$ при $\lambda < 0$. Неявная схема Эйлера *безусловно устойчива*.

3. Схема с весами

$$(u_{i+1} - u_i) / \tau = \lambda(\sigma u_{i+1} + (1 - \sigma)u_i), \quad 0 < \sigma < 1; \quad u_{i+1} = u_i \frac{1 + (1 - \sigma)\tau\lambda}{1 - \sigma\tau\lambda}.$$

Неравенство $1 - \sigma\tau\lambda \leq (1 - \sigma)\tau\lambda \leq 1 - \sigma\tau\lambda$ выполнено при весе $\sigma > 1/2$ при любом шаге τ ; при весе $0 < \sigma < 1/2$ шаг τ должен удовлетворять неравенству $\tau \leq \frac{1}{(\sigma - 1/2)\lambda}$.

4. Схема Рунге – Кутты второго порядка.

Подставляя в схему

$$u_{i+1} = u_i + \frac{\tau}{2} [f(t_i, u_i) + f(t_{i+1}, \tilde{u}_{i+1})], \quad \tilde{u}_{i+1/2} = u_i + \tau f(t_i, u_i)$$

вместо $f = \lambda u$, получаем

$$u_{i+1} = u_i + \frac{\tau}{2} \left[\lambda u_i + \lambda(u_i + \tau\lambda u_i) \right] = u_i \left(1 + \tau\lambda + \frac{\tau^2 \lambda^2}{2} \right), \quad \lambda < 0.$$

Неравенство $(1 + \tau\lambda + \frac{\tau^2 \lambda^2}{2}) < 1$ выполняется, если $\tau \leq 2/|\lambda|$.

5. Неявная схема Адамса третьего порядка

$$(u_{i+1} - u_i)/\tau = \frac{\lambda}{12} (-5u_{i+1} - 8u_i + u_{i-1}), \quad \lambda < 0.$$

Полагая $u_i = q^i$, получим характеристическое уравнение

$$aq^2 + bq + c = 0,$$

где $a = 1 - 5/12\tau\lambda$, $b = -8/12\tau\lambda - 1$, $c = 1/12\tau\lambda$.

Условие (6.36) выполняется при $|q_{1,2}| \leq 1$.

По теореме Виета

$$|q_1 q_2| = \frac{|c|}{a} \leq 1, \quad |q_1 + q_2| = \frac{|b|}{a} \leq 2.$$

Откуда $|c| \leq a$ и $|b| \leq 2a$.

В результате получаем, что схема устойчива, если $\tau \leq 6/|\lambda|$.

6.6. Понятие о жестких задачах

При решении некоторых задач Коши явными методами при увеличении шага происходит катастрофический рост погрешности решения. Обладающие таким свойством задачи получили название *жестких задач*.

Явные методы оказались непригодными для решения жестких задач, т. к. они приводят к большим ограничениям на шаг из-за требований устойчивости в ущерб требованиям точности.

Пример 2. Рассмотрим задачу Коши $y'(t) = -40y(t) + 40\sin t + \cos t$, $y(0) = 1$. Ее решением, как нетрудно проверить, является функция $y(t) = \sin t + e^{-40t}$. На начальном временном отрезке $0 \leq t \leq 0,1$ функция e^{-40t} очень быстро изменяется. Заметим, что, $\lambda = f'_y(t, y(t)) = -40 < 0$, и величина $C(t) = \exp \left\{ \int_{t_0}^t \lambda(\tau) d\tau \right\}$ для оценки погрешности решения с ростом t монотонно убывает.

Если воспользоваться явной схемой Рунге – Кутты второго порядка, то для устойчивости схемы необходимо, чтобы шаг $\tau \leq 2/|\lambda|$ или $\tau \leq 0,05$. При $t > 0,1$ функция $y(t) = \sin t + e^{-40t}$ не обладает быстро меняющейся жесткой компонентой, но попытка увеличить шаг τ при использовании явного метода приводит к неустойчивости решения.

Пример 3. Рассмотрим задачу Коши для системы линейных дифференциальных уравнений:

$$\begin{aligned} y_1'(t) + a_1 y_1 &= 0, \\ y_2'(t) + a_2 y_2 &= 0, \end{aligned} \quad a_2 \gg a_1 > 0, \quad t > 0.$$

Ее решением, как нетрудно проверить, являются функции $y_1(t) = y_1(0)e^{-a_1 t}$, $y_2(t) = y_2(0)e^{-a_2 t}$. Функция $y_2(t)$ очень быстро убывает.

Возьмем явную схему

$$\begin{aligned} (u_{1,i+1} - u_{1,i}) / \tau + a_1 u_{1,i} &= 0, \\ (u_{2,i+1} - u_{2,i}) / \tau + a_2 u_{2,i} &= 0. \end{aligned}$$

Схема устойчива, если выполняются одновременно два условия $\tau \leq 2/a_1$ и $\tau \leq 2/a_2$, что приводит к малому допустимому шагу τ .

Для решения задачи Коши для системы уравнений пригодна неявная схема

$$\begin{aligned}(u_{1,i+1} - u_{1,i}) / \tau + a_1 u_{1,i+1} &= 0, \\ (u_{2,i+1} - u_{2,i}) / \tau + a_2 u_{2,i+1} &= 0,\end{aligned}$$

которая устойчива при любых τ .

При использовании явных методов наличие в решении быстро меняющейся жесткой компоненты даже на том участке, где ее значение пренебрежимо мало, заставляет выбирать шаг из условия абсолютной устойчивости. Применение неявных методов может существенно облегчить решение жесткой задачи.

7. РАЗНОСТНЫЕ МЕТОДЫ РЕШЕНИЯ УРАВНЕНИЯ ТЕПЛОПРОВОДНОСТИ

Процесс распространения тепла в одномерном стержне $0 < x < l$ описывается уравнением теплопроводности

$$c\rho \frac{\partial u}{\partial t} = \frac{\partial}{\partial x} \left(k \frac{\partial u}{\partial x} \right) + f_0(x, t), \quad (7.1)$$

где $u = u(x, t)$ – температура в точке x в момент времени t , c – теплоемкость единицы массы; ρ – плотность; $c\rho$ – теплоемкость единицы длины; k – коэффициент теплопроводности; f_0 – плотность тепловых источников. В общем случае k , c , ρ , f_0 могут зависеть не только от x и t , но и от температуры $u(x, t)$. Если коэффициенты k , c , ρ постоянны, то (7.1) можно записать в виде

$$\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial x^2} + f(x, t),$$

где $a^2 = k / (c\rho)$ – коэффициент температуропроводности.

В дальнейшем без ограничения общности положим, что $a = 1$, $l = 1$.

В самом деле, вводя переменные $x_1 = x/l$, $t_1 = a^2 t / l^2$, $f_1 = l^2 f / a^2$, получим

$$\frac{\partial u}{\partial t_1} = \frac{\partial^2 u}{\partial x_1^2} + f_1(x, t), \quad 0 < x_1 < 1.$$

Будем рассматривать первую краевую задачу в области

$$\Pi = \{0 \leq x \leq 1; 0 \leq t \leq T\}.$$

Требуется найти непрерывное в области Π решение $u = u(x, t)$ задачи

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x, t), \quad (7.2)$$

$$u(x, 0) = u_0(x), \quad 0 \leq x \leq 1; \quad (7.3)$$

$$u(0, t) = \mu_0(t), \quad u(1, t) = \mu_1(t). \quad (7.4)$$

Условие (7.2) характеризует начальное распределение температуры в стержне. Краевые условия (7.3), (7.4) характеризуют распределение температуры на концах стержня.

7.1. Разностные схемы для одномерного уравнения теплопроводности

Введем сетку

$$\omega_{h\tau} = \left\{ \begin{array}{l} (x_i, t_j): x_i = ih, t_j = j\tau, \\ i = 0, 1, 2, \dots, N, h = 1/N, j = 0, 1, \dots, L, \tau = T/L \end{array} \right\}$$

с шагом h по x и с шагом τ по t .

Заменяя вторую производную в (7.2) по x разностным выражением

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_i \approx \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} = L_h(u_i),$$

а первую производную по t разностным отношением

$$\left(\frac{\partial u}{\partial t}\right)_i \approx \frac{u_i(t_{j+1}) - u_i(t_j)}{\tau} = \frac{u_i^{j+1} - u_i^j}{\tau},$$

получим явную схему на 4-точечном шаблоне:

$$\frac{v_i^{j+1} - v_i^j}{\tau} = \frac{v_{i+1}^j - 2v_i^j + v_{i-1}^j}{h^2} + \varphi_i^j, \quad (7.5)$$

$$\begin{array}{ccc} & * & \\ & (x_i, t_{j+1}) & \\ * & * & * \\ (x_{i-1}, t_j) & (x_i, t_j) & (x_{i+1}, t_j). \end{array}$$

Здесь $\varphi_i^j = f(x_i, t_j)$, либо $\varphi_i^j = f(x_i, t_{j+1/2})$, и т. д. Дополнительные условия для определения сеточной функции v :

$$v_i^0 = u_0(x_i), \quad i = 0, 1, 2, \dots, N; \quad (7.6)$$

$$v_0^j = \mu_0(t_j), \quad v_N^j = \mu_1(t_j), \quad j = 0, 1, \dots, L. \quad (7.7)$$

Значения на $(j+1)$ -м временном слое находятся по явной формуле

$$v_i^{j+1} = \left(1 - \frac{2\tau}{h^2}\right)v_i^j + \frac{\tau}{h^2}(v_{i-1}^j + v_{i+1}^j) + \tau\varphi_i^j.$$

Если в разностной схеме $L_h(u_i)$ все значения u_i брать на $(j+1)$ -м временном слое, то получаем полностью *неявную схему* с опережением на 4-х точечном шаблоне:

$$\frac{v_i^{j+1} - v_i^j}{\tau} = \frac{v_{i+1}^{j+1} - 2v_i^{j+1} + v_{i-1}^{j+1}}{h^2} + \varphi_i^j, \quad (7.8)$$

$$\begin{array}{ccc} & * & \\ & (x_{i-1}, t_{j+1}) & (x_i, t_{j+1}) & (x_{i+1}, t_{j+1}) \\ * & * & * \\ & * & \\ & (x_i, t_j). \end{array}$$

Для определения v_i^{j+1} из (7.8) получаем краевую задачу

$$\frac{\tau}{h^2}v_{i-1}^{j+1} - \left(1 + \frac{2\tau}{h^2}\right)v_i^{j+1} + \frac{\tau}{h^2}v_{i+1}^{j+1} = -(v_i^j + \tau\varphi_i^j), \quad i = 1, 2, \dots, N-1;$$

$$v_0^{j+1} = \mu_0(t_{j+1}), \quad v_N^{j+1} = \mu_1(t_{j+1}), \quad j = 0, 1, \dots, L-1,$$

которая на каждом временном слое t_j , $j = 1, 2, \dots, L$ решается методом прогонки.

Если правую часть взять в виде линейной комбинации v_i на $(j+1)$ -м и j -м слоях, то получим разностную схему с весами:

$$\frac{v_i^{j+1} - v_i^j}{\tau} = \sigma L_h(v_i^{j+1}) + (1 - \sigma)L_h(v_i^j) + \varphi_i^j, \quad 0 \leq \sigma \leq 1; \quad (7.9)$$

$$v_i^0 = u_0(x_i), \quad i = 0, 1, 2, \dots, N;$$

$$v_0^j = \mu_0(t_j), \quad v_N^j = \mu_1(t_j), \quad j = 1, \dots, L.$$

Схема (7.9) определена на 6-точечном шаблоне:

$$\begin{array}{ccccc} (x_{i-1}, t_{j+1}) & (x_i, t_{j+1}) & (x_{i+1}, t_{j+1}) & & \\ * & * & * & & \\ * & * & * & & \\ (x_{i-1}, t_j) & (x_i, t_j) & (x_{i+1}, t_j) & & \end{array}$$

В случае $\sigma = 1$ получается чисто неявная схема, а при $\sigma = 0$ – явная схема.

При весе $\sigma = 1/2$ (схема Кранка – Николсона) значения v_i^{j+1} сеточной функции на новом слое определяются из краевой задачи:

$$\frac{\tau}{2h^2} v_{i-1}^{j+1} - \left(1 + \frac{\tau}{h^2}\right) v_i^{j+1} + \frac{\tau}{2h^2} v_{i+1}^{j+1} = -\left(\left(1 - \frac{\tau}{h^2}\right) v_i^j + \frac{\tau}{2h^2} (v_{i-1}^j + v_{i+1}^j) + \tau \varphi_i^j\right);$$

$$i = 1, 2, \dots, N-1;$$

$$v_0^{j+1} = 0, \quad v_N^{j+1} = 0, \quad j = 0, 1, \dots, L-1, \quad L = T/\tau;$$

$$v_i^0 = u(x_i, 0), \quad i = 0, 1, 2, \dots, N.$$

7.2. Устойчивость разностных схем

Рассмотрим устойчивость и сходимость явной схемы. Разностная схема учитывает граничные и начальные условия точно. Рассмотрим задачу определения сеточной функции по явной схеме:

$$v_i^{j+1} = \left(1 - \frac{2\tau}{h^2}\right) v_i^j + \frac{\tau}{h^2} (v_{i-1}^j + v_{i+1}^j) + \tau \varphi_i^j, \quad i = 1, 2, \dots, N-1;$$

$$v_i^0 = u_0(x_i), \quad i = 0, 1, 2, \dots, N; \quad (7.10)$$

$$v_0^{j+1} = 0, \quad v_N^{j+1} = 0, \quad j = 0, 1, \dots, L-1.$$

Если коэффициент при v_i^j неотрицателен, т. е. $\tau \leq \frac{h^2}{2}$, то из задачи (7.10) следует оценка

$$\|\mathbf{v}^{j+1}\|_C \leq \|\mathbf{v}^j\|_C + \tau \|\Phi^j\|_C,$$

где $\|\mathbf{v}\|_C = \max_{0 \leq i \leq N} |v_i|$.

Откуда находим

$$\|\mathbf{v}^{j+1}\|_C \leq \|\mathbf{v}^0\|_C + \sum_{k=0}^{j-1} \tau \|\Phi^k\|_C.$$

Последнее неравенство и выражает устойчивость в сеточной норме C явной схемы по начальным данным и по правой части при условии $\tau \leq \frac{h^2}{2}$, т. е. явная схема условно устойчива.

Для определения v_i^{j+1} из чисто неявной схемы имеем краевую задачу

$$\frac{\tau}{h^2} v_{i-1}^{j+1} - \left(1 + \frac{2\tau}{h^2}\right) v_i^{j+1} + \frac{\tau}{h^2} v_{i+1}^{j+1} = -(v_i^j + \tau \Phi_i^j), \quad i = 1, 2, \dots, N-1 \quad (7.11)$$

с граничными условиями $v_0^{j+1} = 0$, $v_N^{j+1} = 0$, $j = 0, 1, \dots, L-1$, которая решается методом прогонки. Система (7.11) обладает свойством диагонального преобладания (см. 4.2) и верна оценка

$$\|\mathbf{v}^{j+1}\|_C \leq \|\mathbf{v}^j\|_C + \tau \|\Phi^j\|_C,$$

при любых τ и h .

Таким образом, чисто неявная схема безусловно устойчива.

В общем случае схемы с весами $\sigma \geq 1/2$ безусловно устойчивы.

При $0 < \sigma < 1/2$ схема с весами в сеточной норме C условно устойчива, если

$$\tau \leq \frac{h^2}{2(1-\sigma)}, \quad \text{или} \quad \sigma \geq 1 - \frac{h^2}{2\tau}.$$

Введем в рассмотрение погрешность аппроксимации ψ схемы с весами на решении $u = u(x, t)$ (невязку схемы):

$$\psi_i^j = \sigma L_h(u_i^{j+1}) + (1-\sigma)L_h(u_i^j) + \Phi_i^j - \frac{\partial u}{\partial t}, \quad (x, t) \in \omega_{h\tau}.$$

Тогда для погрешности решения $z = u - v$, $(x, t) \in \omega_{h\tau}$ можно получить оценку

$$\|\mathbf{z}^j\|_C \leq 1/2 \sum_{k=0}^{j-1} \tau \|\psi^k\|_C.$$

В общем случае погрешность аппроксимации схемы связана с порядком погрешности аппроксимации производных. Общая схема (7.9) сходится к решению задачи (7.1) в сеточной норме C со скоростью

$$\|\mathbf{u}^j - \mathbf{v}^j\|_C = \|\mathbf{z}^j\|_C = O(\tau + h^2).$$

Если в схеме с весами (7.9) выбрать

$$\sigma = 1/2 \text{ и } \varphi_i^j = \bar{f}_i = f(x_i, t_{j+1/2}), \quad (7.12)$$

то

$$\|\mathbf{u}^j - \mathbf{v}^j\|_C = \|\mathbf{z}^j\|_C = O(\tau^2 + h^2).$$

Если в схеме с весами (7.9) выбрать

$$\begin{aligned} \sigma = \sigma_* = 1/2 - h^2/(12\tau) \text{ и } \varphi_i^j &= \bar{f}_i + \frac{h^2}{12\tau} L_h(\bar{f}_i), \\ L_h(\bar{f}_i) &= \frac{\bar{f}_{i+1} - 2\bar{f}_i + \bar{f}_{i-1}}{h^2}, \quad \bar{f}_i = f(x_i, t_{j+1/2}), \end{aligned} \quad (7.13)$$

то получим схему повышенного порядка аппроксимации:

$$\|\mathbf{u}^j - \mathbf{v}^j\|_C = \|\mathbf{z}^j\|_C = O(\tau^2 + h^4).$$

Свойство асимптотической устойчивости задачи (7.2)–(7.4) при больших значениях t для явной схемы (7.10) и неявной схемы (7.11) совпадают с условиями обычной устойчивости. Однако при весе $\sigma = 1/2$ схема асимптотически устойчива при условии $\tau \leq \frac{h}{\pi}$.

8. МЕТОД КОНЕЧНЫХ ЭЛЕМЕНТОВ

Наряду с методами конечных разностей широко используются проекционные методы Ритца и Галеркина, а точнее их современные варианты – «метод конечных элементов» или «проекционно-сеточные методы». Метод конечных элементов впервые был предложен Р. Курантом в 1943 г. Затем в начале 50-х годов двадцатого столетия специалистами по строительной механике был предложен новый метод решения задач теории упругости. Во многих случаях, когда расчетная область имела сложную геометрию, она разбивалась на подобласти простой геометрии, в каждой из которой решение могло быть найдено аналитически. Эти подобласти получили название конечных элементов, а сам подход – метод конечных элементов. На его основе в настоящее время разработано большое число прикладных пакетов для решения большого круга инженерных и научных задач.

При численном решении краевых задач методом конечных элементов рассматриваемую область разбивают на конечное число элементарных подмножеств стандартной формы (которые и называют *конечными элементами*). Затем используются специально подобранные базисные функции φ_j , которые имеют довольно простой вид (чаще всего – многочлены), отличные от нуля лишь на нескольких соседних элементах.

Многие задачи физики, химии и техники математически описываются с помощью вариационных принципов, при этом решение задачи $u(x)$ является стационарной точкой *вариационного функционала*

$$J(u) = \int_a^b F(x, u, u') dx \quad (8.1)$$

и удовлетворяет краевым условиям:

$$u(a) = u_a, \quad u(b) = u_b, \quad (8.2)$$

где значения u_a, u_b фиксированы.

Как известно из курса вариационного исчисления, непрерывно дифференцируемая функция $u(x)$ должна удовлетворять дифференциальному уравнению

$$-\frac{d}{dx} F'_{u'}(x, u, u') + F'_u(x, u, u') = 0,$$

которое принято называть уравнением Эйлера.

Рассмотрим функционал

$$J(u) = \frac{1}{2} \int_a^b (k(u')^2 + qu^2) dx - \int_a^b u f dx, \quad (8.3)$$

задачи одномерной стационарной диффузии, удовлетворяют краевым условиям (8.2), где $q(x)$, $f(x)$, $k(x)$ – кусочно-непрерывные функции, удовлетворяющие условиям $k(x) \geq k_0 > 0$, $q(x) \geq 0$.

Уравнение Эйлера для задачи одномерной стационарной диффузии принимает вид

$$(k(u'))' + qu = f. \quad (8.4)$$

Можно показать, что функция u является точкой минимума функционала (8.3) тогда и только тогда, когда она является решением краевой задачи (8.4), (8.2).

Приведем решение краевой задачи для одномерной диффузии (8.3), (8.2). Разобьем отрезок $[a, b]$ точками $a = x_0 < x_1 < \dots < x_N = b$ на N элементарных отрезков $[x_{i-1}, x_i]$ длины h_i . В качестве конечного элемента здесь берутся отрезки $[x_{i-1}, x_i]$.

Введем базовые кусочно-непрерывные функции $\varphi_j(x)$ при $j = 1, 2, \dots, N-1$ следующим образом:

$$\varphi_j(x) = \begin{cases} (x - x_{j-1}) / h_j, & \text{при } x \in [x_{j-1}, x_j], \\ (x_{j+1} - x) / h_{j+1}, & \text{при } x \in [x_j, x_{j+1}], \\ 0, & \text{при } x \notin [x_{j-1}, x_{j+1}]. \end{cases} \quad (8.5)$$

Введем функции $\varphi_0(x)$ и $\varphi_N(x)$:

$$\varphi_0(x) = \begin{cases} (x_1 - x) / h_1, & \text{при } x \in [x_0, x_1], \\ 0, & \text{при } x \notin [x_0, x_1]; \end{cases} \quad (8.6)$$

$$\varphi_N(x) = \begin{cases} 0, & \text{при } x \in [x_{N-1}, x_N], \\ (x - x_{N-1}) / h_N, & \text{при } x \notin [x_{N-1}, x_N]. \end{cases}$$

Будем искать приближенное решение задачи в виде

$$u^N(x) = \sum_{j=0}^N \alpha_j \varphi_j(x). \quad (8.7)$$

В силу выбора базовых функций $\varphi_j(x)$ коэффициенты в (8.6) $\alpha_j = u^N(x_j)$, т. к. $\varphi_j(x_j) = 1$ и $\varphi_j(x_i) = 0$ при $x_i \neq x_j$. После введения обозначения $u_j^h = u^N(x_j)$ равенство (8.6) можно переписать в виде

$$u^N(x) = \sum_{j=0}^N u_j^h \varphi_j(x).$$

Согласно методу Ритца приближенные решения $u^N(x)$ определяются как функции, минимизирующие функционал (8.3) на множестве U^N – множестве всех функции вида (8.7), удовлетворяющих краевому условию (8.2).

$$J(u^N) = \min_{v \in V^N} J(v).$$

Согласно необходимому условию экстремума, минимум функции (8.7) достигается при тех значениях α_j , при которых частные производные равны нулю:

$$\frac{\partial}{\partial \alpha_i} J\left(\sum_{j=0}^N \alpha_j \varphi_j(x)\right) = 0. \quad (8.8)$$

Добавляя к этим уравнениям условия (8.2), приходим к системе уравнений, из которой можно определить все значения α_j и тем самым – приближение (8.7). Система (8.8), (8.2) приводится к системе алгебраических уравнений:

$$\sum_{j=1}^N a_{ij} \alpha_j = b_i, \quad i = 1, 2, \dots, N-1; \quad (8.9)$$

$$\alpha_0 = u_a, \quad \alpha_N = u_b, \quad (8.10)$$

$$\text{где } a_{ij} = \int_a^b (k \varphi_i' \varphi_j' + q \varphi_i \varphi_j) dx, \quad b_i = \int_a^b f \varphi_i dx.$$

Базовые функции $\varphi_j(x)$, $\varphi_i(x)$ одновременно могут быть отличными от нуля, если $|i - j| \leq 1$, поэтому при $|i - j| > 1$ коэффициенты a_{ij} в системе (8.9) равны нулю. Таким образом, для определения неизвестных $\alpha_j = u_j^h$ получаем из (8.9) и (8.10) следующую систему уравнений с трехдиагональной матрицей:

$$a_{i,i-1}u_{i-1}^h + a_{i,i}u_i^h + a_{i,i+1}u_{i+1}^h = b_i, \quad i = 1, 2, \dots, N-1; \quad (8.11)$$

$$u_0^h = u_a, \quad u_N^h = u_b. \quad (8.12)$$

Из равенств (8.5) легко видеть

$$\varphi'_j(x) = \begin{cases} 1/h_j, & \text{при } x \in [x_{j-1}, x_j], \\ -1/h_{j+1}, & \text{при } x \in [x_j, x_{j+1}], \\ 0, & \text{при } x \notin [x_{j-1}, x_{j+1}]. \end{cases}$$

Систему уравнений (8.11), (8.12) принято называть системой метода конечных элементов или *проекционно-разностной системой*, которая решается методом прогонки.

В силу свойств выбранных базовых функций $\varphi_j(x)$, коэффициенты a_{ij} и b_i в системе (8.11) можно вычислить по следующим формулам:

$$a_{i,i-1} = -h_i^{-1}k_{i-1/2}^h + h_{i+1/2}q_{i-1/2}^h,$$

$$a_{i,i+1} = -h_{i+1}^{-1}k_{i+1/2}^h + h_{i+1/2}q_{i+1/2}^h, \quad (8.13)$$

$$a_{i,i} = h_i^{-1}k_{i-1/2}^h + h_{i+1}^{-1}k_{i+1/2}^h + h_{i+1/2}q_i^h, \quad b_i = h_{i+1/2}f_i^h,$$

где

$$k_{i-1/2}^h = \frac{1}{h_i} \int_{x_{i-1}}^{x_{i+1}} k(x) dx, \quad q_{i\pm 1/2}^h = \frac{1}{h_{i+1/2}} \int_{x_{i-1}}^{x_{i+1}} q(x) \varphi_i(x) \varphi_{i\pm 1/2}(x) dx,$$

$$q_i^h = \frac{1}{h_{i+1/2}} \int_{x_{i-1}}^{x_{i+1}} q(x) \varphi_i^2(x) dx, \quad f_i^h = \frac{1}{h_{i+1/2}} \int_{x_{i-1}}^{x_{i+1}} f(x) \varphi_i(x) dx.$$

Замечание. При построении системы уравнений метода конечных элементов, как правило, возникает необходимость вычисления определенных интегралов. Эта проблема легко решается применением квадратурных формул.

В заключении отметим, что никакие теоретические положения и советы не могут заменить собственного практического опыта вычислительной работы. Такой опыт можно приобрести, переходя от решения учебных задач к серьезным практическим задачам.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Лапчик, М.П. Численные методы: Учеб.пособие для студ. вузов / М.П.Лапчик, М.И.Рагулина, Е.К.Хеннер. - М.: Академия, 2004. - 384 с.
2. Костомаров, Д.П. Вводные лекции по численным методам /Д.П. Костомаров, А.П.Фаворский.–М.:Логос, 2004.-184с.
3. Калиткин, Н.Н. Численные методы /Н.Н.Калиткин.– М.:Наука, 1978. -510с.
4. Самарский, А.А. Численные методы /А.А.Самарский, А.В.Гулин.–М.: Наука, 1989. - 432с.
5. Самарский, А.А. Введение в численные методы /А.А.Самарский.–М.: Наука, 1987, изд.2-е. - 286с.