

# ВВЕДЕНИЕ В ЧИСЛЕННЫЙ АНАЛИЗ

В. Б. БАРАХНИН  
В. П. ШАПЕЕВ



**В. Б. БАРАХНИН,  
В. П. ШАПЕЕВ**

# **ВВЕДЕНИЕ В ЧИСЛЕННЫЙ АНАЛИЗ**

---

**УЧЕБНОЕ ПОСОВИЕ**



**САНКТ-ПЕТЕРБУРГ · МОСКВА · КРАСНОДАР**  
**2005**

# ОГЛАВЛЕНИЕ

|                                                                                  |    |
|----------------------------------------------------------------------------------|----|
| Предисловие .....                                                                | 6  |
| Глава 1. ПОГРЕШНОСТЬ ЧИСЛЕННОГО РЕШЕНИЯ ...                                      | 7  |
| § 1.1. Классификация погрешностей .....                                          | 7  |
| 1.1.1. Погрешность математической модели .....                                   | 7  |
| 1.1.2. Неустраняемая погрешность .....                                           | 8  |
| 1.1.3. Погрешность метода .....                                                  | 8  |
| 1.1.4. Вычислительная погрешность .....                                          | 8  |
| § 1.2. Оценки погрешности округления .....                                       | 9  |
| 1.2.1. Представление вещественных чисел в ЭВМ ....                               | 9  |
| 1.2.2. Абсолютная и относительная погрешность ...                                | 10 |
| 1.2.3. Погрешности результатов арифметических<br>операций .....                  | 11 |
| 1.2.4. Верные знаки числа .....                                                  | 13 |
| Упражнения .....                                                                 | 14 |
| Глава 2. ИНТЕРПОЛИРОВАНИЕ И ПРИБЛИЖЕНИЕ<br>ФУНКЦИЙ .....                         | 15 |
| § 2.1. Интерполирование алгебраическими<br>многочленами .....                    | 15 |
| 2.1.1. Многочлен Лагранжа .....                                                  | 16 |
| 2.1.2. Многочлен Ньютона .....                                                   | 18 |
| 2.1.3. Оценка погрешности интерполирования ....                                  | 21 |
| 2.1.4. Интерполирование с кратными узлами ....                                   | 23 |
| § 2.2. Численное дифференцирование .....                                         | 25 |
| 2.2.1. Использование интерполяционных формул ....                                | 25 |
| 2.2.2. Метод неопределенных коэффициентов .....                                  | 27 |
| 2.2.3. О вычислительной погрешности формул<br>численного дифференцирования ..... | 30 |

|                                                                |        |
|----------------------------------------------------------------|--------|
| § 2.3. Метод наименьших квадратов .....                        | 32     |
| 2.3.1. Постановка задачи .....                                 | 32     |
| 2.3.2. Выбор базисных функций .....                            | 34     |
| § 2.4. Интерполирование сплайнами .....                        | 36     |
| 2.4.1. Построение кубического сплайна .....                    | 37     |
| 2.4.2. Краевые условия для кубического сплайна ....            | 38     |
| 2.4.3. Сглаживающий сплайн .....                               | 42     |
| Упражнения .....                                               | 47     |
| <br>Глава 3. ЧИСЛЕННОЕ ИНТЕГРИРОВАНИЕ .....                    | <br>49 |
| § 3.1. Квадратурные формулы<br>интерполяционного типа .....    | <br>50 |
| 3.1.1. Формула прямоугольников .....                           | 51     |
| 3.1.2. Формула трапеций .....                                  | 52     |
| 3.1.3. Формула Симпсона .....                                  | 53     |
| 3.1.4. Некоторые обобщения .....                               | 55     |
| 3.1.5. Симметричные формулы .....                              | 56     |
| 3.1.6. Апостериорная оценка погрешности<br>методом Рунге ..... | <br>59 |
| § 3.2. Квадратурные формулы Гаусса .....                       | 61     |
| 3.2.1. Постановка задачи .....                                 | 61     |
| 3.2.2. Существование и единственность<br>формул Гаусса .....   | <br>63 |
| 3.2.3. Свойства квадратурных формул Гаусса .....               | 66     |
| Упражнения .....                                               | 69     |

|                                                             |     |
|-------------------------------------------------------------|-----|
| Глава 4. ПРИБЛИЖЕННЫЕ МЕТОДЫ РЕШЕНИЯ<br>УРАВНЕНИЙ .....     | 71  |
| § 4.1. Постановка задачи .....                              | 71  |
| 4.1.1. Отделение корней. Метод бисекции .....               | 71  |
| § 4.2. Основы теории итерационных методов .....             | 73  |
| 4.2.1. Принцип сжимающих отображений .....                  | 74  |
| 4.2.2. Скорость сходимости метода .....                     | 77  |
| § 4.3. Примеры итерационных методов .....                   | 81  |
| 4.3.1. Метод релаксации .....                               | 82  |
| 4.3.2. Метод Ньютона .....                                  | 84  |
| 4.3.3. Модификации метода Ньютона .....                     | 87  |
| 4.3.4. Методы Чебышева высоких порядков<br>сходимости ..... | 90  |
| 4.3.5. Метод парабол .....                                  | 93  |
| 4.3.6. Метод Эйткена ускорения сходимости .....             | 94  |
| 4.3.7. Сравнение свойств методов .....                      | 97  |
| § 4.4. Решение систем нелинейных уравнений .....            | 98  |
| 4.4.1. Постановка задачи .....                              | 98  |
| 4.4.2. Сходимость стационарных методов .....                | 100 |
| 4.4.3. Примеры итерационных методов .....                   | 103 |
| Упражнения .....                                            | 104 |
| Список литературы .....                                     | 105 |
| Предметный указатель .....                                  | 106 |

## ПРЕДИСЛОВИЕ

Курс численного анализа является тем базисом, на котором зиждется изучение студентами численных методов решения на ЭВМ различных классов математических задач. К сожалению, в программах многих университетов и технических вузов отсутствуют лекции по численному анализу, что приводит к некоторым пробелам в освоении теоретических основ вычислительной математики, наблюдаемым на старших курсах. Практические занятия по этому курсу не могут быть полноценной заменой лекциям, ибо изучая важнейшие факты численного анализа путем решения соответствующих задач, студенты, как правило, «перестают видеть за деревьями лес». Наконец, самостоятельное освоение теории с помощью учебников осложняется тем, что нужные сведения разбросаны по разным книгам. Отметим, что незнание теории может привести к получению неверных результатов даже при расчетах с использованием библиотек стандартных программ.

Данное пособие призвано помочь студентам овладеть основами теории численного анализа, применять их при выполнении заданий по практике на ЭВМ и служить справочником, содержащим необходимые сведения. Пособие содержит важнейшие понятия и снабженные доказательствами теоремы, относящиеся к основным разделам курса: теории погрешностей, интерполированию функций, численному дифференцированию и интегрированию, приближенному решению нелинейных уравнений. В пособии имеется большое количество теоретических задач, а также ряд вычислительных упражнений.

# ПОГРЕШНОСТЬ ЧИСЛЕННОГО РЕШЕНИЯ

## § 1.1. КЛАССИФИКАЦИЯ ПОГРЕШНОСТЕЙ

### 1.1.1. ПОГРЕШНОСТЬ МАТЕМАТИЧЕСКОЙ МОДЕЛИ

При численном решении практических задач всегда следует иметь в виду следующее обстоятельство: никакая математическая модель не может абсолютно точно описать изучаемое явление. В качестве примера рассмотрим процесс колебаний маятника. Простейшее уравнение, описывающее это явление, имеет вид

$$l \frac{d^2 \varphi}{dt^2} + g \varphi = 0,$$

где  $\varphi$  — угол отклонения маятника,  $l$  — его длина,  $g$  — ускорение свободного падения. Из школьного курса физики известно, что это уравнение получено в предположении малости угла  $\varphi$  из уравнения

$$l \frac{d^2 \varphi}{dt^2} + g \sin \varphi = 0.$$

Желание учесть трение приведет нас к уравнению

$$l \frac{d^2 \varphi}{dt^2} + g \sin \varphi + \mu \frac{d\varphi}{dt} = 0, \quad (1.1)$$

которое, в свою очередь, тоже является приближенным, в частности, потому, что в действительности трение зависит от скорости нелинейно. Кроме того, мы не учли вес и растяжение нити, на которой подвешен маятник, силу сопротивления воздуха и т.д. Таким образом, любое математическое описание задачи содержит некоторую погрешность, называемую *погрешностью математической модели*. Уменьшить

эту погрешность можно, как правило, лишь ценой усложнения используемой модели. Более подробно данный вопрос рассматривается в рамках курса «Математическое моделирование».

### 1.1.2. НЕУСТРАНИМАЯ ПОГРЕШНОСТЬ

Выбрав определенную модель, мы неизбежно сталкиваемся с тем, что различные параметры, входящие в уравнения модели, а также начальные данные заданы с некоторой погрешностью, так как они определяются в результате экспериментов и измерений. Так, в рассмотренном примере это касается параметров  $l, g, \mu$ , начального положения маятника и др. Погрешность такого рода называется *неустранимой*, так как она не может быть уменьшена в процессе численного решения. Ее уточнение возможно лишь за счет более точного определения параметров задачи. Заметим, что в некоторых книгах к неустранимой погрешности относят и погрешность математической модели.

### 1.1.3. ПОГРЕШНОСТЬ МЕТОДА

В большинстве случаев исследуемые уравнения, в частности, уравнение (1.1), не могут быть решены аналитически, и для их решения используют тот или иной приближенный численный метод. При этом возникает *погрешность метода*, являющаяся одной из важнейших характеристик любого численного алгоритма. Более подробно о погрешностях различных численных методов будет сказано в соответствующих разделах курса.

### 1.1.4. ВЫЧИСЛИТЕЛЬНАЯ ПОГРЕШНОСТЬ

Так как количество разрядов записи числа на бумаге и в ЭВМ конечно, то при выполнении умножения, деления и других, более сложных, операций, приводящих к увеличению количества разрядов, приходится производить округление результатов, что порождает *вычислительную погрешность* (называемую еще *погрешностью округления*). Способам ее оценки посвящен следующий параграф.



Отметим, что *полная погрешность решения*, то есть разность между реально получаемым и точным решением задачи, не превосходит суммарного значения четырех вышеназванных погрешностей.

## § 1.2. ОЦЕНКИ ПОГРЕШНОСТИ ОКРУГЛЕНИЯ

### 1.2.1. ПРЕДСТАВЛЕНИЕ ДЕЙСТВИТЕЛЬНЫХ ЧИСЕЛ В ЭВМ

В современных ЭВМ приняты две формы записи действительных чисел: с фиксированной точкой и с плавающей точкой (иногда в этих терминах слово «точка» заменяется словом «запятая»).

Представление числа в форме с *фиксированной точкой* означает, что оно записано в *позиционной системе счисления с основанием  $r$* , то есть для числа  $a$ , имеющего вид

$$a = \pm a_n a_{n-1} \dots a_0, a_{-1} a_{-2} \dots,$$

выполняется равенство

$$a = \pm (a_n r^n + a_{n-1} r^{n-1} + \dots + a_0 r^0 + a_{-1} r^{-1} + a_{-2} r^{-2} + \dots).$$

Здесь  $r$  — целое число, большее единицы, числа  $a_i$ , называемые *разрядами*, принадлежат множеству  $\{0, 1, \dots, r-1\}$ .

При  $r = 10$  мы имеем традиционно используемую десятичную систему счисления. В ЭВМ используют, как правило, двоичную систему, когда  $r = 2$ .

Число, записанное с *плавающей точкой*, имеет вид

$$a = M r^p,$$

где  $r$  — основание системы счисления,  $p$  — целое число, называемое *порядком числа  $a$* ,  $M$  — число с фиксированной точкой, называемое *мантиссой числа  $a$* , причем  $r^{-1} \leq |M| < 1$ . Такая нормировка обеспечивает единственность представления любого действительного числа (кроме нуля).

Количество разрядов, отводимое ЭВМ для записи каждого числа, ограничено. Это означает, что подмножество действительных чисел, которым оперирует ЭВМ, является конечным. Наименьшее по модулю машинное число  $M_0$ , отличное от нуля, называется *машинным нулем*. При появлении в

процессе счета числа, меньшего по модулю  $M_0$ , ему присваивается нулевое значение. Наибольшее по модулю машинное число  $M_\infty$  называется *машинной бесконечностью*. При появлении числа, большего  $M_\infty$ , происходит переполнение разрядной сетки, что приводит к *аварийному останову машины (авосту)*. Таким образом, абсолютные значения всех отличных от нуля действительных чисел, представимых данной ЭВМ, лежат между  $M_0$  и  $M_\infty$ .

Если число  $a$  не представимо в ЭВМ точно, оно подвергается *округлению*, которое происходит путем замены числа  $a$  на близкое ему число  $\tilde{a}$ , представимое в ЭВМ точно. Простейшим приемом округления является отбрасывание разрядов мантиссы, выходящих за пределы разрядной сетки. Необходимо уметь оценивать возникающую погрешность округления, а также отслеживать ее изменение в процессе счета.

### 1.2.2. АБСОЛЮТНАЯ И ОТНОСИТЕЛЬНАЯ ПОГРЕШНОСТЬ

Пусть  $x$  — точное значение некоей величины, а  $\tilde{x}$  — некоторое приближение к нему (не обязательно полученное при машинном округлении). Разность

$$\Delta\tilde{x} = \tilde{x} - x$$

называется *абсолютной погрешностью* числа  $\tilde{x}$ . Как правило, эта величина представляет лишь теоретический интерес, ибо точное значение  $x$  обычно неизвестно. Поэтому на практике используют *предельную абсолютную погрешность*  $\Delta(\tilde{x})$ , являющуюся верхней оценкой модуля абсолютной погрешности:  $|\Delta| \leq \Delta(\tilde{x})$ .

Однако абсолютная погрешность не может дать достаточно полного представления о точности вычисления, если не указан сам результат. Бóльшей информативностью обладает отношение

$$\delta\tilde{x} = \frac{|\Delta\tilde{x}|}{|\tilde{x}|},$$

называемое *относительной погрешностью*. Аналогично определяется *предельная относительная погрешность*:

$$\delta(\tilde{x}) = \frac{\Delta(\tilde{x})}{|\tilde{x}|}.$$

Заметим, что относительная погрешность часто выражается в процентах.

Разумеется, если нам известно точное значение величины, то абсолютная и относительная погрешности могут выступать в качестве предельных.

**Пример.** Предположим, что числа с плавающей точкой записываются в ЭВМ в системе счисления с основанием  $r$ , притом для записи мантиссы числа отводится  $k$  разрядов. Пусть число  $a$ , имеющее вид

$$a = \pm r^p \left( \frac{a_1}{r^1} + \frac{a_2}{r^2} + \dots + \frac{a_k}{r^k} + \frac{a_{k+1}}{r^{k+1}} + \dots \right),$$

где  $a_i \in \{0, 1, \dots, r-1\}$ , округляется путем отбрасывания лишних разрядов, в результате чего получается число

$$\tilde{a} = \pm r^p \left( \frac{a_1}{r^1} + \frac{a_2}{r^2} + \dots + \frac{a_k}{r^k} \right).$$

Тогда для предельной абсолютной и предельной относительной погрешностей выполняются равенства:

$$\Delta(\tilde{a}) = r^{p-k}, \quad (1.2)$$

$$\delta(\tilde{a}) = r^{-k+1}. \quad (1.3)$$

**Задача 1.1.** Доказать равенства (1.2) и (1.3).

### 1.2.3. ПОГРЕШНОСТИ РЕЗУЛЬТАТОВ АРИФМЕТИЧЕСКИХ ОПЕРАЦИЙ

Оценим предельную абсолютную погрешность суммы нескольких величин (очевидно, разность может рассматриваться как частный случай суммы). Пусть  $x_1, x_2, \dots, x_n$  — точные значения некоторых величин, а  $\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n$  — их приближенные значения,

$$x = \sum_{i=1}^n x_i, \quad \tilde{x} = \sum_{i=1}^n \tilde{x}_i. \quad (1.4)$$

Тогда

$$|\tilde{x} - x| \leq \sum_{i=1}^n |\tilde{x}_i - x_i| = \sum_{i=1}^n \Delta(\tilde{x}_i).$$

Таким образом, *предельная абсолютная погрешность суммы равна сумме предельных абсолютных погрешностей слагаемых.*

Однако при вычислениях столь быстрый рост погрешности, как правило, не наблюдается, ибо одни приближенные числа округляются с недостатком, а другие с избытком, и их погрешности компенсируют друг друга.

Для предельной относительной погрешности суммы, очевидно, имеет место формула

$$\frac{|\tilde{x} - x|}{|\tilde{x}|} \leq \sum_{i=1}^n \frac{|\tilde{x}_i|}{|\tilde{x}|} \frac{\Delta(\tilde{x}_i)}{|\tilde{x}_i|} = \sum_{i=1}^n \frac{|\tilde{x}_i|}{|\tilde{x}|} \delta(\tilde{x}_i).$$

Отсюда следует, что *предельная относительная погрешность суммы, вообще говоря, может превосходить предельную относительную погрешность каждого из слагаемых.*

**Задача 1.2.** Пусть в формулах (1.4)  $n = 2$ ,  $x_1 = 100$ ,  $\tilde{x}_1 = 105$ ,  $x_2 = -50$ ,  $\tilde{x}_2 = -48$ . Вычислить и сравнить  $\delta(\tilde{x}_1)$ ,  $\delta(\tilde{x}_2)$  и  $\delta(\tilde{x})$ .

**Задача 1.3.** Пусть в формулах (1.4) все  $x_1 > 0$  и все  $\tilde{x}_i > 0$ . Доказать, что

$$\delta(\tilde{x}) \leq \max_i \delta(\tilde{x}_i).$$

Для оценки предельных погрешностей произведения нам потребуется предположение, что абсолютные погрешности значительно меньше самих приближенных величин, то есть значения  $x_i$  и  $\tilde{x}_i$  достаточно близки. Итак, пусть

$$x = \prod_{i=1}^n x_i, \quad \tilde{x} = \prod_{i=1}^n \tilde{x}_i.$$

Тогда, согласно формуле конечных приращений, следующей из формулы Тейлора с остаточным членом в форме Лагранжа,

$$|\tilde{x} - x| = \sum_{i=1}^n |\tilde{x}_i - x_i| \left| \frac{\partial(x_1 \cdot \dots \cdot x_n)}{\partial x_i} \right|_{x_j = \xi_j} = \sum_{i=1}^n |\tilde{x}_i - x_i| \left| \prod_{j=1}^n \frac{\xi_j}{\xi_i} \right|,$$

где  $(\xi_1, \xi_2, \dots, \xi_n)$  — некоторая точка отрезка, соединяющего точки  $(x_1, x_2, \dots, x_n)$  и  $(\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n)$ . Используя сделанное предположение, имеем

$$\Delta(\tilde{x}) \approx \sum_{i=1}^n \frac{|\tilde{x}|}{|\tilde{x}_i|} \Delta(\tilde{x}_i),$$

$$\delta(\tilde{x}) \approx \sum_{i=1}^n \delta(\tilde{x}_i).$$

Таким образом, *предельная относительная погрешность произведения приближенно равна сумме предельных относительных погрешностей сомножителей.*

**Задача 1.4.** Показать, что аналогичное утверждение имеет место и для частного двух величин.

Заметим, что использование формулы конечных приращений часто оказывается полезным при оценке погрешностей достаточно гладких функций с приближенно заданными аргументами.

#### 1.2.4. ВЕРНЫЕ ЗНАКИ ЧИСЛА

Записывая приближенное число, мы обязательно должны указывать погрешность округления. Явная запись вида  $\tilde{x} \pm \Delta(\tilde{x})$  достаточно громоздка, поэтому существует прием, позволяющий судить о погрешности по записи только самого приближенного числа.

Дадим несколько определений.

*Значащими цифрами* числа называются все цифры в его записи, начиная с первой ненулевой слева. Значащая цифра называется *верной*, если единица разряда, соответствующего этой цифре, не меньше абсолютной погрешности числа; в противном случае цифра считается *сомнительной*.

Запись приближенного числа без указания погрешности означает, что все записанные цифры верны.

**Пример.** Запись  $\pi \approx 3,14$  означает, что число  $\pi$  вычислено с погрешностью, не превосходящей 0,01.

**Пример.** Если предельная абсолютная погрешность числа 217 000 не превосходит  $10^3$ , его записывают в виде  $217 \cdot 10^3$ .

Описанный прием дает достаточно грубое представление об истинной погрешности приближенного числа. Поэтому, если приближенное число является промежуточным результатом и должно еще участвовать в последующих вычислениях, следует сохранять одну-две сомнительные цифры.

### УПРАЖНЕНИЯ

1. Радиус круга приближенно равен  $10\text{ см}$ . С какой точностью его надо измерить, чтобы погрешность площади не превышала  $1\text{ см}^2$ ? С какой точностью надо задать число  $\pi$ , чтобы по известной площади круга определить его радиус с погрешностью, меньшей, чем  $0,01$ ?

2. Корни уравнения  $x^2 - 2x + \lg 2 = 0$  нужно получить с четырьмя верными цифрами. С каким числом верных цифр нужно взять свободный член?

## ИНТЕРПОЛИРОВАНИЕ И ПРИБЛИЖЕНИЕ ФУНКЦИЙ

В вычислительной практике часто встречаются функции, значения которых заданы лишь в нескольких точках отрезка. Пусть  $f(x)$  — такая функция,  $f(x_i) = y_i$ ,  $i = 0, 1, \dots, n$ . Возникает задача каким-либо образом доопределить эту функцию в остальных точках отрезка. С этой целью строят функцию  $g(x)$ , совпадающую с  $f(x)$  в точках  $x_i$ . Такой способ выполнения значений функции называется *интерполированием*. Очевидно, в каждом конкретном случае существует бесконечно много вариантов построения функции  $g(x)$ , поэтому к ней предъявляется ряд требований, наиболее естественным из которых является простота вычисления этой функции. Заметим, что к интерполированию нередко прибегают, когда аналитическое выражение для  $f(x)$  известно, но его вычисление слишком трудоемко.

Другой подход к восполнению значений функции  $f(x)$  состоит в построении функции  $h(x)$  так, чтобы отклонение в некоторой норме  $h(x)$  от  $g(x)$  было наименьшим. Этот прием называется *приближением* (*аппроксимацией*), а функция  $h(x)$  — *аппроксимирующей функцией*. Желательно, чтобы она была легко вычисляемой.

### § 2.1. ИНТЕРПОЛИРОВАНИЕ АЛГЕБРАИЧЕСКИМИ МНОГОЧЛЕНАМИ

Рассмотрим задачу интерполирования функции  $f(x)$ , принимающей в точках  $x_0, x_1, \dots, x_n$  (все точки различны; они называются *узлами интерполирования*) значения  $f_0, f_1, \dots, f_n$  соответственно. В качестве интерполирующей функции  $g(x)$  рассмотрим многочлен

$$P_m(x) = \sum_{i=0}^m a_i x^i$$

степени  $m$ , называемый *интерполяционным многочленом*, потребовав, чтобы степень  $P_m(x)$  была наименьшей. По определению,  $P_m(x_j) = f_j$ ,  $j = 0, 1, \dots, n$ . Используя эти равенства для нахождения коэффициентов  $a_i$ , получим систему линейных уравнений с матрицей

$$\begin{pmatrix} 1 & x_0 & \dots & x_0^m \\ 1 & x_1 & \dots & x_1^m \\ \dots & \dots & \dots & \dots \\ 1 & x_n & \dots & x_n^m \end{pmatrix}.$$

При  $m \leq n$  любой ее минор  $(m+1)$ -го порядка есть определитель Вандермонда  $\Delta$ , который в общем случае имеет вид

$$\Delta = \begin{vmatrix} 1 & z_0 & \dots & z_0^k \\ 1 & z_1 & \dots & z_1^k \\ \dots & \dots & \dots & \dots \\ 1 & z_k & \dots & z_k^k \end{vmatrix},$$

причем выполняется равенство

$$\Delta = \prod_{i>j} (z_i - z_j). \quad (2.1)$$

**Задача 2.1.** Доказать равенство (2.1).

Так как в нашем случае все точки  $x_0, x_1, \dots, x_n$  различны, то при  $m < n$  система уравнений для нахождения  $a_i$  не имеет решения, а при  $m = n$  она однозначно разрешима. Таким образом, *интерполяционный многочлен, построенный по  $(n+1)$ -й различной точке, имеет степень не ниже  $n$ , при этом интерполяционный многочлен  $n$ -й степени единствен.*

Изучим различные формы записи этого многочлена.

### 2.1.1. МНОГОЧЛЕН ЛАГРАНЖА

Будем искать интерполяционный многочлен в виде

$$L_n(x) = \sum_{i=0}^n f_i \varphi_i(x),$$

где  $\varphi_i(x)$  — многочлен  $n$ -й степени такой, что



$$\varphi_i(x_j) = \delta_{ij},$$

где

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

— символ Кронекера.

Очевидно,

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

Многочлены  $\varphi_i(x)$  называются *базисными многочленами Лагранжа*.

Итак, интерполяционный многочлен вида

$$L_n(x) = \sum_{i=0}^n f_i \frac{\prod_{j \neq i} (x-x_j)}{\prod_{j \neq i} (x_i-x_j)}$$

называется *интерполяционным многочленом Лагранжа*. Его можно записать в более компактной форме, введя вспомогательный многочлен  $\omega_n(x)$  степени  $n+1$ :

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

Производная этого многочлена в точке  $x_i$  имеет вид

$$\omega'_n(x_i) = (x_i-x_0)(x_i-x_1)\dots(x_i-x_{i-1})(x_i-x_{i+1})\dots(x_i-x_n).$$

Тогда

$$\varphi_i(x) = \frac{\omega_n(x)}{(x-x_i)\omega'_n(x_i)}$$

и, окончательно,

$$L_n(x) = \sum_{i=0}^n f_i \frac{\omega_n(x)}{(x-x_i)\omega'_n(x_i)}.$$

Многочлен Лагранжа имеет особенно простой вид, когда узлы интерполирования являются равноотстоящими, т. е.  $x_i - x_{i-1} = h \equiv \text{const}$ . При этом, если ввести обозначение  $\frac{x-x_0}{h} = t$ , то получим

$$\begin{aligned}\varphi_i(x) &= \frac{\prod_{j \neq i} (x - x_j)}{\prod_{j \neq i} (x_i - x_j)} = \frac{\prod_{j \neq i} (th - jh)}{\prod_{j \neq i} (ih - jh)} = \\ &= \frac{t(t-1) \dots (t-n)}{t-i} \frac{(-1)^{n-i}}{i!(n-i)!} = (-1)^{n-i} \frac{C_n^i}{t-i} \frac{t(t-1) \dots (t-n)}{n!}.\end{aligned}$$

Тогда

$$L_n(x) = L_n(x_0 + th) = (-1)^n \frac{t(t-1) \dots (t-n)}{n!} \sum_{i=0}^n (-1)^i \frac{C_n^i f_i}{t-i}.$$

### 2.1.2. МНОГОЧЛЕН НЬЮТОНА

Интерполяционный многочлен можно рассматривать и как разностный аналог ряда Тейлора, при этом аналогом понятия производной является понятие *разделенной разности* (с точностью до числового коэффициента).

Рассмотрим множество точек  $x_0, x_1, \dots, x_n$  (все точки различны), в которых функция  $f(x)$  соответственно принимает значения  $f_0, f_1, \dots, f_n$ . *Разделенными разностями первого порядка* называются отношения

$$f(x_i, x_{i+1}) = \frac{f_{i+1} - f_i}{x_{i+1} - x_i}, \quad i = 0, 1, \dots, n-1,$$

*разделенными разностями второго порядка* — отношения

$$\begin{aligned}f(x_i, x_{i+1}, x_{i+2}) &= \frac{f(x_{i+1}, x_{i+2}) - f(x_i, x_{i+1})}{x_{i+2} - x_i}, \\ i &= 0, 1, \dots, n-2.\end{aligned}$$

и, вообще, *разделенными разностями  $k$ -го порядка* — отношения

$$\begin{aligned}f(x_i, \dots, x_{i+k}) &= \frac{f(x_{i+1}, \dots, x_{i+k}) - f(x_i, \dots, x_{i+k-1})}{x_{i+k} - x_i}, \\ i &= 0, 1, \dots, n-k.\end{aligned}$$

**Лемма.** *Имеет место равенство*

$$f(x_i, \dots, x_{i+k}) = \sum_{j=i}^{i+k} \frac{f(x_j)}{\prod_{\substack{l \neq j \\ l=i}}^{i+k} (x_j - x_l)}.$$

**Доказательство** проведем по индукции. При  $k = 1$  утверждение сразу следует из определения разделенных разностей первого порядка. Пусть утверждение доказано для  $k \leq m$ . Тогда

$$\begin{aligned} f(x_i, \dots, x_{i+m+1}) &= \frac{f(x_{i+1}, \dots, x_{i+m+1}) - f(x_i, \dots, x_{i+m})}{x_{i+m+1} - x_i} = \\ &= \frac{1}{x_{i+m+1} - x_i} \left( \sum_{j=i+1}^{i+m+1} \frac{f(x_j)}{\prod_{\substack{l \neq j \\ l=i+1}}^{i+m+1} (x_j - x_l)} - \sum_{j=i}^{i+m} \frac{f(x_j)}{\prod_{\substack{l \neq j \\ l=i}}^{i+m} (x_j - x_l)} \right). \end{aligned}$$

В полученном выражении  $f(x_i)$  и  $f(x_{i+m+1})$  встречаются по одному разу, притом соответствующие коэффициенты имеют требуемый вид. Остальные  $f(x_j)$  входят в выражение дважды, коэффициент при  $f(x_j)$  равен

$$\begin{aligned} &\frac{1}{\prod_{\substack{l \neq j \\ l=i+1}}^{i+m+1} (x_j - x_l)} - \frac{1}{\prod_{\substack{l \neq j \\ l=i}}^{i+m} (x_j - x_l)} = \\ &= \frac{(x_j - x_i) - (x_j - x_{i+m+1})}{(x_{i+m+1} - x_i) \prod_{\substack{l \neq j \\ l=i}}^{i+m+1} (x_j - x_l)} = \frac{1}{\prod_{\substack{l \neq j \\ l=i}}^{i+m+1} (x_j - x_l)}, \end{aligned}$$

что и требовалось установить.

**Следствие 1.** *Разделенная разность является линейным оператором функции  $f$ .*

**Следствие 2.** *Разделенная разность есть симметрическая функция своих аргументов.*

Перейдем к выводу формулы интерполяционного многочлена Ньютона. Для этого запишем многочлен Лагранжа в виде

$$L_n(x) = L_0(x) + \sum_{j=1}^n (L_j(x) - L_{j-1}(x)),$$

где  $L_j(x)$  — многочлен Лагранжа, построенный по узлам  $x_0, x_1, \dots, x_j$ , причем  $L_0(x) = f(x_0)$ . Заметим, что  $L_j(x) - L_{j-1}(x)$  есть многочлен степени  $j$ , обращающийся в нуль в точках  $x_0, x_1, \dots, x_{j-1}$ , то есть

$$L_j(x) - L_{j-1}(x) = A_j(x - x_0)(x - x_1) \dots (x - x_{j-1}).$$

Так как  $L_j(x_j) = f(x_j)$ , то

$$A_j = \frac{f(x_j) - L_{j-1}(x_j)}{(x_j - x_0)(x_j - x_1) \dots (x_j - x_{j-1})}.$$

Отсюда, пользуясь определением многочлена Лагранжа, получим

$$\begin{aligned} A_j &= \frac{1}{\prod_{k=0}^{j-1} (x_j - x_k)} \left( f(x_j) - \sum_{i=0}^{j-1} f(x_i) \frac{\prod_{k \neq i} (x_j - x_k)}{\prod_{k \neq i} (x_i - x_k)} \right) = \\ &= \sum_{i=0}^j \frac{f(x_i)}{\prod_{l \neq i} (x_i - x_l)} = f(x_0, \dots, x_j). \end{aligned}$$

Таким образом,

$$\begin{aligned} L_n(x) &= f(x_0) + (x - x_0)f(x_0, x_1) + (x - x_0)(x - x_1)f(x_0, x_1, x_2) + \dots \\ &\quad \dots + (x - x_0)(x - x_1) \dots (x - x_{n-1})f(x_0, x_1, \dots, x_n). \end{aligned}$$

Интерполяционный многочлен, записанный в таком виде, называется *интерполяционным многочленом Ньютона*.

Отметим, что многочлены Лагранжа и Ньютона представляют собой всего лишь различные формы записи одного и того же интерполяционного многочлена, при этом многочлен Лагранжа целесообразно использовать в тех случаях, когда интерполируется несколько функций, значения которых заданы в одних и тех же узлах, а многочлен Ньютона — при интерполировании одной и той же функции с постепенно увеличивающимся числом узлов.

### 2.1.3. ОЦЕНКА ПОГРЕШНОСТИ ИНТЕРПОЛИРОВАНИЯ

Замена функции  $f(x)$  интерполяционным многочленом  $L_n(x)$  приводит к появлению погрешности

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

называемой также *остаточным членом интерполяционной формулы*. Очевидно, в узлах погрешность равна нулю. Для оценки остаточного члена в иных точках рассмотрим вспомогательную функцию

$$\varphi(s) = f(s) - L_n(s) - K\omega_n(s),$$

причем постоянную  $K$  выберем из условия  $\varphi(x) = 0$ , где  $x$  — точка, в которой оценивается погрешность. Для этого зафиксируем точку  $x$  и положим

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

Таким образом, функция  $\varphi(s)$  обращается в нуль по крайней мере в  $n+2$  точках:  $x_0, x_1, \dots, x_n, x$ . Пусть функция  $f(x)$  имеет  $n+1$  непрерывную производную. Согласно теореме Ролля производная  $\varphi'(s)$  обращается в нуль по крайней мере в  $n+1$  точке,  $\varphi''(s)$  — в  $n$  точках и т. д., наконец,  $\varphi^{(n+1)}(s)$  обращается в нуль по крайней мере в 1 точке  $\xi$ , причем эта точка принадлежит отрезку  $[y_1, y_2]$ , где  $y_1 = \min(x_0, x_1, \dots, x_n, x)$ ,  $y_2 = \max(x_0, x_1, \dots, x_n, x)$ . Так как

$$\varphi^{(n+1)}(s) = f^{(n+1)}(s) - K(n+1)!,$$

то из условия  $\varphi^{(n+1)}(\xi) = 0$  получим

$$K = \frac{f^{(n+1)}(\xi)}{(n+1)!},$$

откуда

$$f(x) - L_n(x) = \frac{f^{(n+1)}(\xi)\omega_n(x)}{(n+1)!}, \quad \xi \in [y_1, y_2].$$

Окончательно имеем

$$|f(x) - L_n(x)| \leq \frac{M_{(n+1)}(\xi)}{(n+1)!} |\omega_n(x)|, \quad (2.2)$$

где  $M_{n+1} = \max_{x \in [y_1, y_2]} |f^{(n+1)}(x)|$ . Отсюда видно, что погрешность в точке  $x$  относительно невелика, если

$$\min(x_0, x_1, \dots, x_n) < x < \max(x_0, x_1, \dots, x_n).$$

В противном случае, называемом *экстраполированием*, резко возрастает значение  $|\omega_n(x)|$ .

При  $x_i - x_{i-1} = h \equiv \text{const}$  можно говорить о *порядке сходимости* погрешности к нулю при  $h$ , стремящемся к нулю. Этот порядок равен степени полинома  $\omega_n(x)$  в остаточном члене, то есть  $n+1$ , поскольку в него будет входить множителем  $h^{n+1}$ .

С помощью разделенных разностей можно получить другую оценку погрешности интерполирования, не используя предположения о существовании производных высокого порядка функции  $f(x)$ :

$$\begin{aligned} r_n(x) &= f(x) - L_n(x) = f(x) - \sum_{i=0}^n f_i \prod_{j \neq i} \frac{(x - x_j)}{(x_i - x_j)} = \\ &= \prod_{i=0}^n (x - x_i) \left( \frac{f(x)}{\prod_{i=0}^n (x - x_i)} + \sum_{j=0}^n \frac{f(x_j)}{(x_j - x) \prod_{i \neq j} (x_i - x_j)} \right) = \\ &= \omega_n(x) f(x, x_0, x_1, \dots, x_n). \end{aligned} \quad (2.3)$$

Сравнивая оценки (2.2) и (2.3), видим, что при  $f \in C^{n+1}$  существует точка  $\xi \in [y_1, y_2]$ , для которой

$$f(x, x_0, x_1, \dots, x_n) = \frac{f^{n+1}(\xi)}{(n+1)!}.$$

Если же  $f \notin C^{n+1}$ , то оценка (2.2) не выполняется, и порядок сходимости погрешности в общем случае будет меньше, чем  $n+1$ .

#### 2.1.4. ИНТЕРПОЛИРОВАНИЕ С КРАТНЫМИ УЗЛАМИ

Рассмотрим общую постановку задачи интерполирования.

В узлах  $x_k$ ,  $k = 0, 1, \dots, m$ , среди которых нет совпадающих, заданы значения функции  $f(x_k)$  и ее производных  $f^{(i)}(x_k)$ ,  $i = 1, \dots, N_k - 1$ , при этом число  $N_k$  называется *кратностью узла*  $x_k$ . Иными словами, в каждом узле интерполирования известны значения

$$f(x_k), f'(x_k), \dots, f^{(N_k-1)}(x_k),$$

то есть всего известно  $N_0 + N_1 + \dots + N_m$  величин. Требуется построить многочлен  $H_n(x)$  степени  $n = N_0 + N_1 + \dots + N_m - 1$ , удовлетворяющий условиям

$$H_n^{(i)}(x_k) = f^{(i)}(x_k), \quad k = 0, 1, \dots, m, \quad i = 0, 1, \dots, N_k - 1. \quad (2.4)$$

Многочлен  $H_n(x)$  называется *интерполяционным многочленом Эрмита*.

Докажем, что интерполяционный многочлен Эрмита существует и единствен. Пусть

$$H_n(x) = \sum_{i=0}^n a_i x^i.$$

Подставляя это выражение в формулы (2.4), получим систему линейных уравнений относительно  $a_i$ , при этом число уравнений равно числу неизвестных и составляет  $N_0 + N_1 + \dots + N_m$ . Рассматривая условия (2.4) с нулевой правой частью, получим

$$H_n^{(i)}(x_k) = 0, \quad k = 0, 1, \dots, m, \quad i = 0, 1, \dots, N_k - 1, \quad (2.5)$$

откуда следует, что  $x_k$  является корнем кратности  $N_k$  многочлена  $H_n(x)$ . Таким образом, многочлен  $H_n(x)$  степени  $n$  имеет с учетом кратности не менее  $N_0 + N_1 + \dots + N_m = n + 1$  корня. Это означает, что  $H_n(x) \equiv 0$ , то есть все коэффициенты  $a_i$  равны нулю. Итак, однородная система (2.5) имеет только нулевое решение, вследствие чего неоднородная система (2.4) однозначно разрешима при любых правых частях, причем коэффициенты  $a_i$  линейно выражаются через значения  $f^{(i)}(x_k)$ . Тем самым многочлен Эрмита можно представить в виде

$$H_n(x) = \sum_{k=0}^m \sum_{i=0}^{N_k-1} c_{ki}(x) f^{(i)}(x_k),$$

где  $c_{ki}(x)$  — многочлены степени  $n$ . Явный вид этих многочленов мы не приводим ввиду их громоздкости.

**Задача 2.2.** Доказать, что многочлен Эрмита, построенный по значениям функции  $f_i$  и ее производной  $f'_i$ , которые заданы в узлах  $x_i$ ,  $i = 0, 1, \dots, m$ , имеет вид

$$H(x) = \sum_{i=0}^m \left( f_i + (-2\varphi'_i(x_i)f_i + f'_i)(x - x_i) \right) \varphi_i^2(x),$$

где  $\varphi_i(x)$  — базисные многочлены Лагранжа.

Оценка остаточного члена интерполирования проводится так же, как для многочлена Лагранжа. Введем вспомогательную функцию

$$\varphi(s) = f(s) - H_n(s) - K\omega(s),$$

где  $\omega(s) = \prod_{i=0}^m (s - x_i)^{N_i}$ . Постоянную  $K$  выберем из условия  $\varphi(x) = 0$ , где  $x$  — точка, в которой оценивается погрешность:

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

Функция  $\varphi(s)$  имеет с учетом кратности не менее  $N_0 + N_1 + \dots + N_m + 1 = n + 1$  корня. Последовательно применяя теорему Ролля к производным функции  $\varphi$  и учитывая кратность корней, получим, что существует точка  $\xi$  такая, что  $\varphi^{(n+1)}(\xi) = 0$ , откуда



$$f(x) - H_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^m (x - x_i)^{N_i}. \quad (2.6)$$

## § 2.2. ЧИСЛЕННОЕ ДИФФЕРЕНЦИРОВАНИЕ

### 2.2.1. ИСПОЛЬЗОВАНИЕ ИНТЕРПОЛЯЦИОННЫХ ФОРМУЛ

*Численным дифференцированием* называется процесс вычисления приближенного значения производной  $f^{(k)}(x)$  некоторой функции  $f(x)$  по ее значениям  $f_0, f_1, \dots, f_n$ , заданным в точках  $x_0, x_1, \dots, x_n$ .

Простейшие формулы численного дифференцирования получаются путем дифференцирования интерполяционных формул. Пусть задана *сетка* (т.е. набор узлов)  $x_0 < x_1 < \dots < x_n$  с шагами  $h_i = x_i - x_{i-1}$ ,  $i = 1, 2, \dots, n$ . Рассмотрим многочлен Лагранжа  $L_{1,i}(x)$ , построенный для функции  $f(x)$  по точкам  $x_{i-1}$  и  $x_i$ . Очевидно,

$$L'_{1,i}(x) = \frac{f_i - f_{i-1}}{h_i}.$$

Это выражение можно принять за приближенное значение  $f'(x)$  в любой точке  $x \in [x_{i-1}, x_i]$ .

В практических задачах часто встречается равномерная сетка,  $h_i = h$ ,  $i = 1, 2, \dots, n$ . Последовательно полагая  $x = x_i$  и  $x_{i-1}$  и увеличивая во втором случае все индексы на 1, мы получим следующие важные формулы:

$$f'(x_i) \approx \frac{f_i - f_{i-1}}{h} \text{ — разность назад, обозначаемая } f_{x,i}^-;$$

$$f'(x_i) \approx \frac{f_{i+1} - f_i}{h} \text{ — разность вперед, обозначаемая } f_{x,i}^+.$$

Используем для получения формул численного дифференцирования многочлен Лагранжа  $L_{2,i}(x)$ , построенный для функции  $f(x)$  по трем точкам  $x_{i-1}, x_i, x_{i+1}$ . Найдем его производную:

$$\begin{aligned} L'_{2,i}(x) &= \frac{2x - x_i - x_{i+1}}{h_i(h_i + h_{i+1})} f_{i-1} - \\ &- \frac{2x - x_{i-1} - x_{i+1}}{h_i h_{i+1}} f_i + \frac{2x - x_{i-1} - x_i}{h_{i+1}(h_i + h_{i+1})} f_{i+1}. \end{aligned}$$

Считая сетку равномерной и полагая  $x = x_i$ , получим формулу *центральной разности*

$$f'(x_i) \approx \frac{f_{i+1} - f_{i-1}}{2h},$$

обозначаемой  $f_{x,i}^\circ$ ; полагая  $x = x_{i-1}$  и  $x = x_{i+1}$ , после сдвига индексов получим соответственно формулы

$$f'(x_i) \approx \frac{-3f_i + 4f_{i+1} - f_{i+2}}{2h}, \quad (2.7)$$

$$f'(x_i) \approx \frac{f_{i-2} - 4f_{i-1} + 3f_i}{2h}. \quad (2.8)$$

Вычислим погрешность полученных формул. Пусть  $f \in C^3$  на соответствующем интервале. Разлагая  $f(x_{i+1})$  и  $f(x_{i-1})$  в ряд Тейлора с остаточным членом в форме Лагранжа, после приведения подобных имеем

$$|f_{x,i}^\circ - f'(x_i)| = \left| \frac{f_{i+1} - f_{i-1}}{2h} - f'(x_i) \right| = \left| \frac{h^2}{6} f'''(\xi^{(1)}) \right| \leq M_3 \frac{h^2}{6}.$$

Здесь и далее в этом параграфе  $M_n = \sup_{(x_{i-1}, x_{i+1})} |f^{(n)}(x)|$ ,  $\xi^{(1)}$  есть некоторая точка интервала  $(x_{i-1}, x_{i+1})$ .

Аналогичным образом, предполагая  $f \in C^2$ , получим оценки:

$$|f_{x,i}^- - f'(x_i)| = \left| \frac{h}{2} f''(\xi^{(2)}) \right| \leq M_2 \frac{h}{2},$$

$$|f_{x,i}^+ - f'(x_i)| = \left| \frac{h}{2} f''(\xi^{(3)}) \right| \leq M_2 \frac{h}{2}.$$

Заметим, что если погрешность какой-либо приближенной формулы на равномерной сетке с шагом  $h$  является величиной  $O(h^k)$ , то эту формулу называют *формулой  $k$ -го порядка*.

**Задача 2.3.** Оценить погрешность формул (2.7) и (2.8). Сравнить результат с оценкой погрешности  $f_{x,i}^+$  и  $f_{x,i}^-$ .

Получим приближенное значение для  $f''(x)$ , вычисляя вторую производную многочлена  $L_{2,i}(x)$ :

$$L''_{2,i}(x) = \frac{2}{h_i(h_i + h_{i+1})}f_{i-1} - \frac{2}{h_i h_{i+1}}f_i + \frac{2}{h_{i+1}(h_i + h_{i+1})}f_{i+1}.$$

Правая часть этого равенства не зависит от  $x$ . В частности, на равномерной сетке имеем

$$f_{-\frac{+}{x},i} = \frac{f_{\frac{+}{x}} - f_{\frac{-}{x}}}{h} = \frac{f_{i-1} - 2f_i + f_{i+1}}{h^2}.$$

Оценим погрешность последней формулы:

$$|f_{-\frac{+}{x},i} - f'(x_i)| = \left| \frac{h^2}{12} f^{IV}(\xi) \right| \leq M_4 \frac{h^2}{12}.$$

**Задача 2.4.** Оценить погрешность формул, задаваемых разностными отношениями  $f_{-\frac{-}{x},i}$  и  $f_{\frac{+}{x},i}$ .

Очевидно, для приближенного вычисления производных более высокого порядка требуется использовать многочлены Лагранжа больших степеней, увеличивая количество узлов, участвующих в аппроксимации. Кроме того, нельзя заранее предсказать точность получаемой формулы. Поэтому на практике обычно используют другие способы получения формул численного дифференцирования, один из которых будет рассмотрен ниже.

### 2.2.2. МЕТОД НЕОПРЕДЕЛЕННЫХ КОЭФФИЦИЕНТОВ

Суть метода неопределенных коэффициентов заключается в следующем. Выражение для производной  $k$ -го порядка в некоторой точке  $x = x_*$  ищется в виде линейной комбинации заданных значений функции в узлах  $x_0, x_1, \dots, x_n$ :

$$f^{(k)}(x_*) = \sum_{i=0}^n c_i f(x_i) + R(f), \quad (2.9)$$

где  $R(f)$  — остаточный член, зависящий от функции. Коэффициенты  $c_i$  подбираются из условия  $R(f) = 0$ , когда  $f = 1, x, x^2, \dots, x^n$ . В результате для нахождения  $c_i$  получаем систему  $n + 1$  линейных алгебраических уравнений:

Эта система однозначно разрешима, так как ее определитель есть определитель Вандермонда, при этом для получения ненулевого решения необходимо и достаточно выполнения неравенства  $n > k$ .

$$\left| \sum_{i=0}^n \tilde{c}_i y_i^j = (y^j)^{(k)} \right|_{y=y_*}, \quad j = 0, 1, \dots, n,$$
$$\sum_{i=0}^n \tilde{c}_i (x_i - d)^j = ((x - d)^j)^{(k)} \Big|_{x=x_*}, \quad j = 0, 1, \dots, n.$$

Таким образом, при вычислении значений  $s_i$  можно брать в качестве  $x_*$  точку, наиболее удобную для выкладок. Обычно полагают  $x_* = 0$ .

Поясним смысл требования, предъявляемого к коэффициентам  $c_i$  в формуле (2.9), согласно которому на любом многочлене степени не выше  $n$  должно выполняться условие  $R(f) = 0$ . Заменяем в правой части формулы (2.9) функцию  $f$  ее разложением по формуле Тейлора в окрестности точки  $x_*$ . Тогда степень членов ряда Тейлора, содержащихся в остатке  $R(f)$ , будет не меньше  $n + 1$ . В частности, на равномерной сетке с шагом  $h$ , имеющей  $n + 1$  узлов, метод неопределенных коэффициентов аппроксимирует  $k$ -ю производную достаточно гладкой функции с порядком  $n + 1 - k$ .

**Задача 2.5.** Доказать последнее утверждение.

При некоторых дополнительных предположениях метод неопределенных коэффициентов может дать в тех же условиях формулы более высокого порядка точности. Пусть узлы равномерной сетки расположены симметрично относительно точки  $x_*$ , то есть  $x_i - x_* = x_* - x_{n-i}$ . Не уменьшая общности, далее считаем  $x_* = 0$ , откуда  $x_i = -x_{n-i}$ . Используя последнее равенство при  $0 \leq i \leq \frac{n}{2}$ , получим систему для определения  $c_i$ :

$$\begin{aligned} c_0(-x_n)^j + c_1(-x_{n-1})^j + \dots + c_{n-1}x_{n-1}^j + c_nx_n^j = \\ = \delta_{kj}k!, \quad j = 0, 1, \dots, n, \end{aligned} \quad (2.11)$$

где  $\delta_{kj}$  — символ Кронекера. Осуществим переобозначение неизвестных  $c_i = \tilde{c}_{n-i}$ ,  $i = 0, 1, \dots, n$ , и домножим  $j$ -е уравнение на число  $(-1)^j$ ,  $j = 0, 1, \dots, n$ . Тогда система (2.11) примет вид

$$\begin{aligned} \tilde{c}_0(-1)^jx_n^j + \tilde{c}_1(-1)^jx_{n-1}^j + \dots + \tilde{c}_{n-1}(-1)^j(-x_{n-1})^j + \\ + \tilde{c}_n(-1)^j(-x_n)^j = \delta_{kj}(-1)^jk!, \quad j = 0, 1, \dots, n. \end{aligned} \quad (2.12)$$

Очевидно, системы (2.11) и (2.12) имеют одинаковые матрицы, а вектора их правых частей при четном  $k$  совпадают, то есть  $c_i = \tilde{c}_i = c_{n-i}$  а при нечетном  $k$  различаются лишь знаком, то есть  $c_i = -\tilde{c}_i = -c_{n-i}$  ( $j = 0, 1, \dots, n$ ). Таким образом, на равномерной сетке, симметричной относительно точки  $x_*$ , формула (2.9) при четном  $k$  симметрична, а при нечетном  $k$  — антисимметрична.

Далее, если в сделанных предположениях  $k$  четно, а число узлов нечетно (т. е.  $n$  четно), то после подстановки в формулу (2.9) разложения в ряд Тейлора функции  $f$  произойдет сокращение членов порядка  $n + 1$ , в результате чего производная будет аппроксимирована с порядком  $n + 2 - k$ . Такой же порядок аппроксимации достигается при нечетном  $k$  и четном числе узлов (нечетном  $n$ ). Сказанное означает, что подсчет производной в средней точке может привести к повышению порядка аппроксимации.

Подробная сводка формул численного дифференцирования, полученных методом неопределенных коэффициентов, приведена, например, в [4].

### 2.2.3. О ВЫЧИСЛИТЕЛЬНОЙ ПОГРЕШНОСТИ ФОРМУЛ ЧИСЛЕННОГО ДИФФЕРЕНЦИРОВАНИЯ

При использовании формул численного дифференцирования следует иметь в виду, что уменьшение шага сетки  $h$ , осуществляемое с целью уменьшения погрешности метода, может привести к росту влияния вычислительной погрешности и, как следствие, к увеличению полной погрешности.

Напомним, что при аппроксимации первой производной разностью назад имеет место следующая оценка:

$$|f_{x,i}^- - f'(x_i)| = \left| \frac{f_i - f_{i-1}}{h} - f'(x_i) \right| = \left| \frac{h}{2} f''(\xi) \right| \leq M_2 \frac{h}{2}.$$

Наличие вычислительной погрешности приводит к тому, что вместо точных значений  $f_i$ ,  $f_{i-1}$  мы имеем дело с приближенными значениями  $\tilde{f}_i = f_i + \delta_i$ ,  $\tilde{f}_{i-1} = f_{i-1} + \delta_{i-1}$ . Таким образом, вычислительная погрешность разностной производной окажется равной  $\delta_{x,i}^- = (\delta_i - \delta_{i-1})/h$ .

Пусть для погрешностей  $\delta_i$ ,  $\delta_{i-1}$  имеет место оценка  $|\delta_i| \leq \delta$ ,  $|\delta_{i-1}| \leq \delta$ . Тогда

$$|\delta_{x,i}^-| \leq 2\delta/h.$$

Значит, для полной погрешности  $r$  имеем оценку

$$|r| \leq g(h) = M_2 h/2 + 2\delta/h.$$

Нетрудно убедиться, что минимум функции  $g(h)$  достигается при

$$h = h_0 = 2\sqrt{\delta/M_2}$$

и составляет

$$g(h_0) = 2\sqrt{\delta M_2},$$

причем при  $h < h_0$  значение  $g(h)$  резко возрастает (см. рис. 2.1).

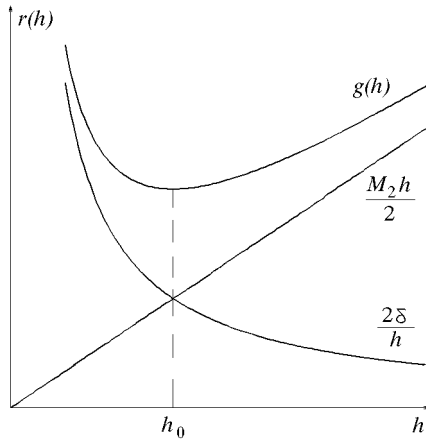


Рис. 2.1. Погрешность численного дифференцирования

Итак, в рассмотренном примере полная погрешность имеет порядок  $O(\sqrt{\delta})$ . Так как величина  $\delta$  не может быть меньше точности представления чисел в ЭВМ, то значение  $f_{x,i}^-$  вычисляется в лучшем случае с половиной верных разрядов. Это свидетельствует о некорректности рассмотренного здесь способа численного дифференцирования при наличии погрешности в значениях  $f_i$ . Этот способ нуждается в регуляризации [5].

Заметим, что при вычислении производных более высокого порядка наличие в знаменателе разностного отношения величины  $h^k$ ,  $k > 1$ , приведет к усилению влияния вычислительной погрешности. Так, при подсчете четвертой производной со вторым порядком точности минимум полной погрешности достигается, если  $h_0 = O(\delta^{1/6})$ . Пусть  $\delta \approx 10^{-12}$ , тогда  $h_0 \approx 0,01$ , и, следовательно, погрешность метода и вычислительная погрешность имеют порядок  $10^{-4}$ .

## § 2.3. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

### 2.3.1. ПОСТАНОВКА ЗАДАЧИ

Если в задаче приближенного восстановления функции  $f(x)$  дискретно заданные на некотором интервале  $(a, b)$  значения  $f(x_i)$ ,  $i = 0, 1, \dots, n$  — входные данные — имеют погрешности  $\varepsilon_i$ , превышающие по величине допустимую погрешность интерполирования, то решение с помощью интерполяционных многочленов, принимающих эти значения в узлах, не может считаться удовлетворительным, так как оно будет неприемлемо низкой точности. Здесь во многих случаях могут быть полезны методы приближения функций, которые удовлетворяют другим требованиям близости к  $f(x)$  искомого приближения  $\varphi(x)$ . В частности, если погрешность входных данных носит случайный характер, то вместо алгоритмов интерполирования хорошие результаты решения задачи могут дать такие способы приближения функций, как *метод наименьших квадратов*, алгоритмы построения сглаживающих сплайнов. Если заведомо известно, что решение должно обладать определенной гладкостью, то  $\varphi(x)$  можно брать из соответствующего класса функций. Требование равенства  $\varphi(x_i) = f(x_i)$  можно заменить требованием близости функций в некоторой норме, так что случайные отклонения от некоторой гладкой функции в таких методах будут сглажены в решении, и максимум отклонения  $\varphi(x)$  от  $f(x)$  на всем интервале может оказаться меньше погрешности  $\varepsilon$  входных данных задачи. Здесь и далее в этой главе через  $\varepsilon$  будет обозначаться некоторое среднее значение погрешностей  $\varepsilon_i$ ,  $i = 0, 1, \dots, n$ .

Пусть имеются функция  $f(x)$  и множество функций  $\varphi(x)$ , принадлежащих линейному нормированному пространству функций. В задаче аппроксимации требуется для заданных  $f(x)$  и  $\varepsilon$  найти такую  $\Phi(x)$ , чтобы выполнялось неравенство  $\|f(x) - \Phi(x)\| \leq \varepsilon$ . В задаче наилучшего приближения надо найти функцию  $\Phi(x)$ , удовлетворяющую соотношению

$$\|f(x) - \Phi(x)\| = \inf_{\varphi} \|f(x) - \varphi(x)\| = \nu. \quad (2.13)$$

Существование и единственность наилучшего приближения зависят от выбора пространства и множества  $\varphi(x)$ .



В методе наименьших квадратов рассматривается гильбертово пространство  $L_2(\rho)$  действительных функций, интегрируемых с квадратом с весом  $\rho(x) > 0$  на  $[a, b]$ . Норма в нем равна  $\|f\|_{L_2} = \sqrt{(f, f)}$  со скалярным произведением

$$(f, \varphi) = \int_a^b \rho(x) f(x) \varphi(x) dx. \quad (2.14)$$

Аппроксимирующую функцию берем в виде линейной комбинации

$$\Phi(x) = \sum_{j=0}^k c_j \varphi_j(x). \quad (2.15)$$

Подставив ее в (2.13), получим

$$\|f(x) - \Phi(x)\|_{L_2}^2 = (f, f) - 2 \sum_{j,m=0}^k c_j (f, \varphi_j) + \sum_{j=0}^k c_j c_m (\varphi_j, \varphi_m) = \nu. \quad (2.16)$$

Приравнивая нулю производные по коэффициентам  $c_j$  от левой части (необходимое условие минимума квадратичной относительно переменных  $c_j$  функции (2.16)), получим систему линейных уравнений

$$\sum_{j=0}^k (\varphi_m, \varphi_j) c_j = (f, \varphi_m), \quad 0 \leq m \leq k. \quad (2.17)$$

Матрица системы (2.17) — матрица Грама — симметричная и положительно определенная. Ее определитель в случае линейно независимых функций  $\varphi_j$  отличен от нуля. Следовательно, для такого множества базисных функций наилучшее среднеквадратичное приближение существует и единственно. Чтобы приближенно численно решить систему (2.17), введем скалярное произведение на конечном множестве точек

$$(f, \varphi) = \sum_{i=0}^n \rho_i f(x_i) \varphi(x_i), \quad \rho > 0, \quad (2.18)$$

где  $n+1$  — число всех узлов на рассматриваемом интервале  $[a, b]$  с известными значениями  $f(x_i)$ .

### 2.3.2. ВЫБОР БАЗИСНЫХ ФУНКЦИЙ

Выбор конкретного множества функций  $\varphi_j(x)$  зависит от таких свойств функции  $f(x)$ , как периодичность, симметричность, наличие асимптотики, экспоненциальный или логарифмический вид и т. д. Для решения этой задачи используются алгебраические многочлены, тригонометрические полиномы, множества из базисных функций ортогональных систем.

Если известна информация о точности задания  $f(x_i)$ , то в этом случае весу  $\rho_i$  придают смысл точности данных в конкретной точке  $x_i$ : чем выше точность  $\varepsilon_i$  заданного значения  $f_i$  в этой точке в сравнении с точностью в других точках, тем большее значение веса  $\rho_i$  приписывается ей. Обычно значение веса пропорционально  $\varepsilon_i^{-2}$ .

В качестве первого множества базисных функций возьмем  $\varphi_j = x^j$ . Уравнения (2.17) примут вид

$$\sum_{j=0}^k (x^m, x^j) c_j = (f, x^m), \quad 0 \leq m \leq k, \quad (2.19)$$

$$(x^m, x^j) = \sum_{i=0}^n \rho_i x_i^{m+j}, \quad (f, x^m) = \sum_{i=0}^n \rho_i f_i x_i^m.$$

Данная система функций неортогональна, поэтому при больших  $k$  задача (2.19) плохо обусловлена. При  $k$  порядка 10 она обусловлена удовлетворительно и зачастую решается без затруднений. При плохой обусловленности задачи (2.19) используются базисы, состоящие из ортогональных многочленов Чебышева, Лежандра, Лагерра и др. Свойство ортогональности заключается в том, что для каждого из указанных типов многочленов существует отрезок  $[a, b]$ , на котором обращаются в нуль скалярные произведения многочленов разного порядка с весовой функцией  $\rho(x)$ :

$$\int_a^b \rho(x) \varphi_j(x) \varphi_m(x) dx = 0, \quad j \neq m.$$

Для наиболее гладкой аппроксимации функций, с минимальным числом и амплитудой отклонений, берутся многочлены

Чебышева  $T_j(x)$ , которые определены на отрезке  $[-1, 1]$  и ортогональны на нем с весовой функцией  $\rho(x) = (1 - x^2)^{-1/2}$ . Их значения при подсчете скалярных произведений удобно находить по рекуррентным формулам

$$\begin{aligned} T_{j+1}(x) &= 2x T_j(x) - T_{j-1}(x), \\ T_0(x) &= 1, \quad T_1(x) = x. \end{aligned} \quad (2.20)$$

Многочлены Лежандра ортогональны на отрезке  $[-1, 1]$  с весовой функцией  $\rho(x) = 1$  и могут быть вычислены по рекуррентным формулам

$$\begin{aligned} L_{j+1}(x) &= [(2j+1)xL_j(x) - jL_{j-1}(x)]/(j+1), \\ L_0(x) &= 1, \quad L_1(x) = x. \end{aligned} \quad (2.21)$$

Простота условий ортогональности имеет некоторые достоинства.

Для того чтобы построить аппроксимирующую функцию  $\Phi(x)$  с помощью ортогональных многочленов (2.20), (2.21), удобно заданный отрезок  $[a, b]$  с помощью линейного преобразования  $\bar{x} = (2x - a - b)/(b - a)$  перевести в отрезок  $[-1, 1]$ , а после решения задачи для данных  $f(\bar{x}_i)$ ,  $i = 0, 1, \dots, n$  с помощью замены  $x = (a + b - (a - b)\bar{x})/2$  выписать решение на интервале  $[a, b]$ . Использование базиса из ортогональных полиномов позволяет получить хорошо обусловленную систему (2.17) при больших  $k$ .

При  $k = n$  среднеквадратичная аппроксимация совпадает с лагранжевой интерполяцией. Как уже отмечалось, наличие значительной случайной погрешности во входных данных требует отказаться от интерполирования. Хорошее сглаживание следует ожидать при  $k \ll n$ . Но если  $k$  слишком мало, то для удовлетворительного описания сложного вида  $f(x)$  может не хватить малого количества выбранных функций (числа коэффициентов  $c_j$ ) в представлении  $\Phi(x)$ . Их оптимальное число зависит от функции  $f(x)$ , числа узлов  $n$ , их расположения, весов и от выбранной системы  $\varphi_j(x)$ .

Если известна величина погрешности входных данных  $\varepsilon$ , то оптимальное число  $k$  определяют следующим образом. При некотором выбранном  $k$  из условия (2.16) находят  $c_j^{(k)}$ ,  $1 \leq j \leq k$ , вычисляют полученное при этом *среднеквадратичное уклонение*

$$\delta_k = \left( \sum_{i=0}^n \frac{(\Phi(x_i) - f_i)^2}{n+1} \right)^{1/2}.$$

Если  $\delta_k \gg \varepsilon$ , т. е. погрешность аппроксимации много больше погрешности входных данных, то количества коэффициентов недостаточно для описания  $f(x)$ , и надо увеличить  $k$ . Если  $\delta_k \ll \varepsilon$ , то старшие коэффициенты аппроксимации недостоверны, и надо уменьшить  $k$ . Если же  $\delta_k \approx \varepsilon$ , то число коэффициентов оптимально.

Расчет можно начать с  $k = 1$ , когда заведомо  $\delta \gg \varepsilon$ , и постепенно увеличивать число коэффициентов до тех пор, пока не выполнится условие  $\delta_k \approx \varepsilon$ . Если при этом  $k \ll n$ , то вид аппроксимирующей функции, т. е. множество  $\varphi_j(x)$ , выбран удачно. Если же  $k_{\text{опт}} \sim n$ , то необходимо подобрать более подходящую аппроксимирующую функцию. При небольшом  $n$  выбор оптимального  $k$  является весьма трудной задачей, поэтому в таком случае особенно важен удачный выбор множества  $\varphi_j(x)$ .

## § 2.4. ИНТЕРПОЛИРОВАНИЕ СПЛАЙНАМИ

Построение интерполяционного многочлена Лагранжа или Ньютона с использованием большого числа узлов интерполирование на отрезке  $[a, b]$  может привести к плохому приближению интерполируемой функции из-за возрастания вычислительной погрешности. Кроме того, построенный многочлен будет иметь высокую степень, что тоже весьма нежелательно. Этих неприятностей можно избежать, разбив отрезок  $[a, b]$  на частичные отрезки и построив на каждом из них многочлен невысокой степени, так или иначе приближенный к заданной функции  $f(x)$ . Иногда такой прием называют *кусочно-полиномиальным интерполированием*.

Итак, пусть в попарно различных узлах некоторой сетки  $x_i$ ,  $i = 0, 1, \dots, n$  заданы значения  $f_i$  функции  $f(x)$  скалярного аргумента  $x$ . Решение задачи построения функции  $S(x)$ , которая в узлах равна  $f_i$ , принадлежит классу  $C^k$  ( $k \geq 0$ ) на  $x_0 < x < x_n$  и на каждом интервале  $x_{i-1} < x < x_i$  имеет различное представление  $S_i$ , называется *сплайновым интерполированием*, а сама функция  $S(x)$  — *сплайном*. Сплаины в силу хороших аппроксимационных свойств, относительной

простоты и экономичности их построения нашли многочисленные приложения.

Во многих приложениях в качестве сплайна берут многочлен  $S(x) = \sum_{j=0}^{k+1} a_j x^j$ . Его коэффициенты в общем случае различны на различных интервалах, но удовлетворяют условиям согласования — непрерывности  $S(x)$  и  $k$  ее производных в узлах. Чаще всего берут многочлены первой или третьей степени. В последнем случае сплайн называется *кубическим*.

#### 2.4.1. ПОСТРОЕНИЕ КУБИЧЕСКОГО СПЛАЙНА

Опишем процесс построения кубического сплайна. На отдельном интервале  $[x_{i-1}, x_i]$  многочлен  $S(x)$  можно взять в виде

$$S_i(x) = \sum_{j=0}^3 a_{ij}(x - x_{i-1})^j, \quad i = 1, 2, \dots, n. \quad (2.22)$$

Нетрудно видеть, что сформулированная задача пока недоопределена. В задаче имеется  $4n$  неизвестных  $a_{ij}$  — по 4 коэффициента на  $n$  интервалах. А из требования выполнения условий задачи имеем  $4n - 2$  соотношений:  $2n$  соотношений

$$S_i(x_i) = f_i, \quad S_i(x_{i-1}) = f_{i-1}, \quad i = 1, 2, \dots, n \quad (2.23)$$

и  $2n - 2$  соотношений

$$S'_i(x_i) = S'_{i+1}(x_i), \quad i = 1, 2, \dots, n - 1, \quad (2.24)$$

$$S''_i(x_i) = S''_{i+1}(x_i), \quad i = 1, 2, \dots, n - 1. \quad (2.25)$$

Введем обозначения  $M_i = S''(x_i)$ ,  $h_i = x_i - x_{i-1}$ . Поскольку  $S''(x)$  — линейная функция на  $[x_{i-1}, x_i]$ , ее вид определяется двумя значениями:

$$S''_i(x) = M_{i-1} \frac{x_i - x}{h_i} + M_i \frac{x - x_{i-1}}{h_i}, \quad x_{i-1} \leq x \leq x_i. \quad (2.26)$$

Очевидно, (2.26) удовлетворяет условию (2.25). Проинтегрируем (2.25) дважды по  $x$  и, требуя выполнения условий (2.23), получим на отрезке  $[x_{i-1}, x_i]$  равенство

$$S_i(x) = M_{i-1} \frac{(x_i - x)^3}{6h_i} + M_i \frac{(x - x_{i-1})^3}{6h_i} + \\ + (f_{i-1} - \frac{M_{i-1}h_i^2}{6}) \frac{(x_i - x)}{h_i} + (f_i - \frac{M_i h_i^2}{6}) \frac{(x - x_{i-1})}{h_i}, \quad (2.27) \\ i = 1, 2, \dots, n.$$

Дифференцирование (2.27) дает

$$S'_i(x) = -M_{i-1} \frac{(x_i - x)^2}{2h_i} + M_i \frac{(x - x_{i-1})^2}{2h_i} - (f_{i-1} - \frac{M_{i-1}h_i^2}{6}) \frac{1}{h_i} + \\ + (f_i - \frac{M_i h_i^2}{6}) \frac{1}{h_i}, \quad i = 1, 2, \dots, n. \quad (2.28)$$

Подставив (2.28) в (2.24), после простых преобразований получим соотношение

$$\frac{1}{6}(h_i M_{i-1} + 2(h_i + h_{i+1})M_i + h_{i+1}M_{i+1}) = \frac{\nabla f_{i+1}}{h_{i+1}} - \frac{\nabla f_i}{h_i}, \quad (2.29)$$

где

$$\nabla f_{i+1} = f_{i+1} - f_i, \quad i = 1, 2, \dots, n-1,$$

которое означает, что осредненное значение (по трем соседним значениям) второй производной сплайна в узле  $x_i$  равняется второй разностной производной аппроксимируемой функции в этом узле. Все условия сформулированной выше задачи будут выполнены.

Соотношения (2.29) представляют собой систему из  $n-1$  линейных алгебраических уравнений относительно  $n+1$  неизвестных  $M_i$ . Для однозначной разрешимости задачи систему (2.29) необходимо доопределить. Рассмотрим наиболее употребительные способы доопределения задачи — краевые условия.

#### 2.4.2. КРАЕВЫЕ УСЛОВИЯ ДЛЯ КУБИЧЕСКОГО СПЛАЙНА

Если известны значения  $f''(x_0)$  и  $f''(x_n)$ , полагаем

$$M_0 = f''(x_0), \quad M_n = f''(x_n). \quad (2.30)$$

Укажем два простых и эффективных приближенных способа реализации условий (2.30), основанных на их полиномиальной и разностной аппроксимации, которые не требуют дополнительной входной информации в задаче, кроме заданных  $f_i$  ( $i = 0, 1, \dots, n$ ). Если значения  $f''(x_0)$ ,  $f''(x_n)$  неизвестны или известны с недостаточной точностью, то на левом краю строим интерполяционный многочлен третьей степени  $P_0$ , совпадающий с  $f(x)$  в точках  $x_0, x_1, x_2, x_3$ , на правом краю —  $P_n$ , совпадающий с  $f(x)$  в точках  $x_n, x_{n-1}, x_{n-2}, x_{n-3}$ . В краевые условия вместо точных значений производных от интерполируемой функции  $f''(x_0)$  и  $f''(x_n)$  подставим вычисленные значения  $P_0''(x_0)$ ,  $P_n''(x_n)$ . Почти такой же результат по точности получится, если в (2.30) подставить на левом краю вместо  $f''(x_0)$  разностную вторую производную

$$2\left(\frac{f_0}{h_1(h_1 + h_2)} - \frac{f_1}{h_1 h_2} + \frac{f_2}{h_2(h_1 + h_2)}\right)$$

и на правом краю вместо  $f''(x_n)$  разностную вторую производную

$$2\left(\frac{f_{n-2}}{h_{n-1}(h_{n-1} + h_n)} - \frac{f_{n-1}}{h_{n-1} h_n} + \frac{f_n}{h_n(h_{n-1} + h_n)}\right).$$

Когда интерполируемая функция имеет достаточную гладкость, точность интерполирования во многих случаях улучшится, если на краю построить интерполяционный многочлен четвертой степени по значениям  $f(x)$  в пяти крайних точках отрезка и использовать его в краевых условиях. Такую же точность даст замена в (2.30) производных от функции вторыми разностными производными, построенными с помощью односторонних разностей по пяти крайним точкам отрезка. Правда, существенного увеличения точности по сравнению с предыдущими способами реализации краевых условий в этом случае зачастую не наблюдается.

Следует отметить, что способы с аппроксимацией повышенной точности условий (2.30) в численных экспериментах с данными достаточно гладких функций по точности не уступают способам, перечисленным ниже, и несущественно уступают только случаю, когда в краевых условиях берутся точные значения производных от  $f(x)$ .

Если известны значения  $f'(x_0)$  и  $f'(x_n)$ , то, рассматривая (2.28) для первого интервала и полагая  $i = 1$ ,  $x = x_0$ ,  $S'(x_0) = f'(x_0)$ , имеем

$$\frac{h_1}{3}M_0 + \frac{h_1}{6}M_1 = \frac{\nabla f_1}{h_1} - f'(x_0). \quad (2.31)$$

Аналогично, используя на правом краю значение сплайна в точке  $x = x_n$ , получим

$$\frac{h_n}{6}M_{n-1} + \frac{h_n}{3}M_n = f'(x_n) - \frac{\nabla f_n}{h_n}. \quad (2.32)$$

Когда значения  $f'(x_0)$  и  $f'(x_n)$  в формулах (2.31), (2.32) известны с недостаточной точностью, их, как и в первом случае, лучше заменить на значения  $P'_0(x_0)$  и  $P'_n(x_n)$ . При этом формулы (2.31), (2.32) можно преобразовать к виду

$$M_0 + \frac{M_1}{2} = P''_0(x_0) + \frac{P''_0(x_1)}{2}. \quad (2.33)$$

На правом краю аналогично получим

$$\frac{M_{n-1}}{2} + M_n = P''_n(x_n) + \frac{P''_n(x_{n-1})}{2}. \quad (2.34)$$

Как и в первом способе, здесь эффективно применение разностной аппроксимации производных от функции  $f(x)$  в краевых условиях.

Наконец, в качестве краевых условий можно потребовать непрерывность третьей производной сплайна  $S'''_i = \frac{M_i - M_{i-1}}{h_i}$  в точках  $x = x_1, x = x_{n-1}$ . Приравнивая третью производную на первом отрезке (слева в точке  $x_1$ ) и на втором отрезке (справа в точке  $x_1$ ), находим

$$\frac{M_0}{h_1} - \frac{h_2 + h_1}{h_2 h_1} M_1 + \frac{M_2}{h_2} = 0. \quad (2.35)$$

Аналогичное соотношение

$$\frac{M_{n-2}}{h_{n-1}} - \frac{h_n + h_{n-1}}{h_n h_{n-1}} M_{n-1} + \frac{M_n}{h_n} = 0 \quad (2.36)$$

имеет место на правом краю. Формулы (2.35) и (2.36) представляют собой линейную экстраполяцию функции  $M(x)$ . Выразим  $M_0$  из (2.29) при  $i = 1$  и исключим его из (2.35), результат запишем в виде



$$M_1 = \frac{h_1 - h_2}{h_1 + 2h_2} M_2 + \frac{6(h_1 \nabla f_2 - h_2 \nabla f_1)}{h_1(h_1 + h_2)(h_1 + 2h_2)}. \quad (2.37)$$

Аналогично, исключая из (2.36)  $M_n$  с помощью (2.29) при  $i = n - 1$ , получим соотношение между  $M_{n-1}$  и  $M_{n-2}$ :

$$M_{n-1} = \frac{h_n - h_{n-1}}{(h_n + 2h_{n-1})} M_{n-2} + \frac{6(h_{n-1} \nabla f_n - h_n \nabla f_{n-1})}{h_n(h_n + h_{n-1})(h_n + 2h_{n-1})}. \quad (2.38)$$

Соотношения (2.37), (2.38) имеют особо простой вид в случае  $h_1 = h_2$  и  $h_{n-1} = h_n$  соответственно:  $M_1 = \delta^2 f_1 / h_1^2$  и  $M_{n-1} = \delta^2 f_{n-1} / h_{n-1}^2$ , где  $\delta^2 f_i = \nabla f_{i+1} - \nabla f_i$ .

В первом и втором способах доопределения задачи к системе  $n - 1$  уравнений (2.29) присоединяются недостающие два уравнения — краевые условия. В третьем способе на первом этапе берется система, составленная из  $n - 3$  уравнений (2.29) с индексами  $i = 2, 3, \dots, n - 2$  и дополненная уравнениями (2.37) и (2.38). Во всех трех случаях получим систему линейных алгебраических уравнений относительно  $M_i$  с трехдиагональной матрицей, имеющей диагональное преобладание. Для решения этой системы имеется экономичный по числу арифметических действий (порядка  $8n$ ) алгоритм прогонки (прямое исключение по Гауссу), устойчивый в случае матрицы с диагональным преобладанием. В первых двух случаях прогонкой находим сразу все  $M_i$  ( $i = 0, 1, \dots, n$ ), в третьем случае на первом этапе находим  $M_i$  только для индексов  $i = 1, 2, \dots, n - 1$ , а неизвестные  $M_0$  и  $M_n$  вычислим на втором этапе из (2.35) и (2.36) соответственно (проектруем  $M(x)$ ). Задача решена однозначно.

Отметим, что третий способ доопределения задачи дает точность, уступающую первым двум только в случаях точного задания производных от  $f(x)$  в краевых условиях или при их аппроксимации с повышенным порядком. Таким образом, третий способ может применяться на практике во многих случаях.

Естественно, могут быть случаи комбинаций из перечисленных типов краевых условий. Например, на левом краю задано условие первого типа, на правом краю — условие третьего типа. В силу очевидности реализации этих случаев здесь их отдельно не обсуждаем.

После того как найдены все  $M_i$ , значения  $S_i(x)$  на

$[x_{i-1}, x_i]$ , принимаемые за приближенные значения  $f(x)$ , могут быть вычислены по формуле (2.27). Приближенные значения  $f'(x)$  вычисляются по формуле (2.28), а значения  $f''(x)$  — по формуле (2.26).

Перечисленные способы доопределения задачи позволяют строить сплайны, решающие задачу интерполирования для гладких данных с высокой точностью. Так например, в численных экспериментах интерполирования по заданным с равномерным шагом  $h = 0,1$  на отрезке  $[0, 1]$  значениям таких функций, как  $\sin x$ ,  $\exp x$  и других гладких функций, которые вместе со своими производными имеют значения порядка 1, при использовании первых двух типов краевых условий с точными значениями производных от этих функций на краях отрезка погрешность интерполирования на крайних интервалах ( $i = 1, 10$ ) имеет величину порядка  $10^{-7}$ , убывая по величине при переходе от крайних интервалов к внутренним ( $i = 5, 6$ ). На неравномерной сетке можно добиться выравнивания погрешностей на всех интервалах и улучшить результаты.

Для данных от функции, имеющей на интервале  $(x_0, x_1)$  достаточного порядка непрерывные производные, при стремлении  $\max_{i=1,n}(h_i)$  к нулю результат интерполирования с помощью кубического сплайна сходится с четвертым порядком. Например, это можно наблюдать в численных экспериментах с реализацией краевых условий первым способом при точном задании на краях вторых производных интерполируемой функции или при их аппроксимации с повышенным порядком точности. Для достижения нужной точности результата может потребоваться провести вычисления всех величин в задаче с помощью арифметики высокой точности. При вычислениях с 16 двоичными разрядами и  $h < 0,1$  зачастую погрешность метода сравнима с погрешностью округления чисел на ЭВМ. Первая производная от сплайна приближает первую производную от  $f(x) = 0$  на порядок хуже, чем сплайн — саму функцию, а вторая производная приближает  $f''(x)$  еще на один порядок грубее.

Отметим также, что условия вида  $f'(x_0) = 0$ ,  $f''(x_0) = 0$  ( $f'(x_n) = 0$ ,  $f''(x_n) = 0$ ), используемые при решении краевых задач для упругого гибкого стержня, являясь точными в отдельных случаях закрепления стержня, не подходят в общем

случае и дают значительную погрешность в решении при замене на них краевых условий. Таким образом, в общем случае грубо заменять нулями производные от  $f(x)$  при реализации краевых условий в рассматриваемой задаче.

### 2.4.3. СГЛАЖИВАЮЩИЙ СПЛАЙН

С помощью сплайнов можно решить и задачу об приближении некоторой функции  $f(x)$ , значения которой  $f_i$  в узлах сетки  $x_i$ ,  $i = 0, 1, \dots, n$  заданы с некоторой погрешностью  $\varepsilon$ . В этом случае целесообразно, чтобы приближающая функция не совпадала в узлах с  $f_i$ , а в некоторой окрестности узлов имела меньшую кривизну, чем приближаемая, т. е. была более гладкой и в среднем лучше приближала  $f(x)$  на интервале  $[x_0, x_n]$ . Как отмечено в § 2.3, точность приближения в решении этой задачи зависит от входных данных: вида  $f(x)$ , числа  $n$ , расположения узлов, величины погрешности  $\varepsilon_i$ , является ли погрешность случайной или систематической.

Для решения этой задачи можно использовать кубический сплайн. Здесь отметим его одно важное свойство [8]: интерполирующий кубический сплайн  $S(x)$  — единственная функция, которая при интерполировании  $f_i$  дает минимум функционала  $\int_{x_0}^{x_n} [u''(x)]^2 dx$  в классе  $W_2^2[a, b]$ , состоящем из функций  $u(x)$ , имеющих на  $[x_0, x_n]$  суммируемые с квадратом вторые производные и удовлетворяющих конкретным краевым условиям, рассмотренным при решении задачи построения интерполирующего сплайна. При краевых условиях  $S''(x_0) = 0$ ,  $S''(x_n) = 0$  функционал  $\int_{x_0}^{x_n} [u''(x)]^2 dx$  при  $u(x) = S(x)$  достигает минимума на всем классе функций  $W_2^2[a, b]$ . Если заметить, что вторая производная определяет кривизну линии, то геометрический смысл данного экстремального свойства сплайна очевиден: среди всех функций, интерполирующих дискретно заданные функции  $f_i$ ,  $i = 0, 1, \dots, n$ , он обладает в некоторой норме минимумом осредненной кривизны на интервале  $[x_0, x_n]$ . Потенциальная энергия изгиба упругого стержня пропорциональна его кривизне. Это означает, что упругий стержень, шарнирно закрепленный в точках  $(x_i, f_i)$  (при дополнительных краевых условиях, разобранных

в предыдущих пунктах), обладает минимумом потенциальной энергии изгиба.

Потребуем, чтобы искомая сглаживающая функция  $S(x)$  минимизировала на классе  $W_2^2[x_0, x_n]$  функционал

$$\Psi(u) = \int_{x_0}^{x_n} [u''(x)]^2 dx + \sum_{i=0}^n \rho_i [u(x_i) - f_i]^2, \quad (2.39)$$

где  $\rho_i$  — некоторые положительные числа. В функционале  $\Psi(u)$  скомбинированы интерполяционные условия прохождения искомой кривой вблизи заданных точек и условие минимальности осредненной кривизны. Чем больше весовые коэффициенты  $\rho_i$ , тем больший вклад в функционал вносят интерполяционные условия, тем ближе к заданным точкам  $(x_i, f_i)$  проходит сглаживающая кривая и тем больше у нее осредненная кривизна.

Можно показать, что минимум функционала  $\Psi(u)$  также достигается на некотором кубическом сплайне, называемом *сглаживающим сплайном*. Отсюда возникает следующий подход к решению задачи: пусть искомый сглаживающий сплайн, как и интерполирующий, определяется своими значениями  $\mu_i$  (которые в общем случае отличаются от  $f_i$ ), и на нем функционал (2.39) достигает своего минимума. Поскольку кубический сплайн, согласно (2.29), определяется множеством своих значений  $\mu_i$ ,  $i = 0, 1, \dots, n$  в узлах  $x_i$ ,  $i = 0, 1, \dots, n$ , то минимизация  $\Psi(u)$  сводится к нахождению минимума функции от переменных  $\mu_0, \mu_1, \dots, \mu_n$ .

Рассмотрим алгоритм построения сглаживающего сплайна в случае  $M_0 = 0$ ,  $M_n = 0$ . Согласно предыдущему на частичном интервале  $[x_{i-1}, x_i]$  имеет место (2.26). С учетом этого (2.39) можно записать в виде

$$\Psi(S) = \sum_{i=1}^n \int_{x_{i-1}}^{x_i} \left[ M_{i-1} \frac{x_i - x}{h_i} + M_i \frac{x - x_{i-1}}{h_i} \right]^2 dx + \sum_{i=0}^n \rho_i (\mu_i - f_i)^2.$$

Вычисляя интеграл в последней формуле, получим

$$\begin{aligned} & \sum_{i=1}^n \int_{x_{i-1}}^{x_i} \left[ M_{i-1} \frac{x_i - x}{h_i} + M_i \frac{x - x_{i-1}}{h_i} \right]^2 dx = \\ & = \sum_{i=1}^{n-1} M_i \left[ \frac{h_i}{6} M_{i-1} + \frac{h_i + h_{i+1}}{3} M_i + \frac{h_{i+1}}{6} M_{i+1} \right] = (AM, M). \end{aligned}$$

Здесь квадратная трехдиагональная матрица  $A = \{a_{ij}\}$  с элементами  $a_{i,i-1} = \frac{h_i}{6}$ ,  $a_{i,i} = \frac{h_i + h_{i+1}}{3}$ ,  $a_{i,i+1} = \frac{h_{i+1}}{6}$ ,  $i = 1, 2, \dots, n-1$  симметрична, имеет строгое диагональное преобладание и, следовательно, положительно определена. Согласно (2.29)  $M$  линейно выражается через  $\mu = (\mu_0, \mu_1, \dots, \mu_n)$ , поэтому  $\Psi(S)$  — положительно определенная форма от  $\mu$ . В качестве экстремума у нее может быть только минимум. Соотношения

$$\frac{\partial \Phi_1}{\partial \mu_s} \equiv \frac{\partial}{\partial \mu_s}(Am, m) + 2\rho_s(\mu_s - f_s) = 0, \quad s = 0, 1, \dots, n$$

— необходимые условия ее экстремума, где вектор  $m$  определяется из (2.29) через значения  $S(x_i) = \mu_i$ ,  $i = 0, 1, \dots, n$ . Систему (2.29) можно записать в векторной форме:

$$Am = H\mu \quad (2.40)$$

с прямоугольной матрицей  $H$  размера  $(n-1) \times (n+1)$ , у которой элементы  $h_{ii} = \frac{1}{h_i}$ ,  $h_{i,i+1} = (-\frac{1}{h_i} - \frac{1}{h_{i+1}})$ ,  $h_{i,i+2} = \frac{1}{h_{i+1}}$ ,  $i = 1, 2, \dots, n-1$ , а остальные элементы равны нулю. Матрица  $A$  не зависит от  $\mu$ . Поэтому, используя (2.40), получим

$$\begin{aligned} \frac{\partial}{\partial \mu_s}(Am, m) &= 2\left(\frac{\partial(Am)}{\partial \mu_s}, m\right) = 2\left(\frac{\partial(H\mu)}{\partial \mu_s}, m\right) = \\ &= 2\left(\frac{\partial \mu}{\partial \mu_s}, H^T m\right) = 2(H^T m)_s, \end{aligned}$$

где верхний индекс  $T$  означает транспонирование матрицы. Последнее равенство позволяет записать условие минимума в виде

$$H^T m + P\mu = Pf, \quad (2.41)$$

где  $P$  — диагональная матрица с элементами  $p_{ii} = \rho_i$ ,  $i = 0, 1, \dots, n$ . Умножая (2.41) на  $HP^{-1}$ , получим

$$HP^{-1}H^T m + H\mu = Hf.$$

Из последнего уравнения и (2.40) следует

$$(A + HP^{-1}H^T)m = Hf. \quad (2.42)$$

Матрица системы (2.42) пятидиагональна, симметрична и положительно определена. Для ее решения существует много хороших методов. Для средних размеров эффективен метод квадратного корня. После того как найден  $m$ , вектор точных значений сглаживающего сплайна  $\mu$  определяется по формуле

$$\mu = f - P^{-1}H^T m, \quad (2.43)$$

которая следует из (2.41). При вычислении значений сплайна по формуле (2.27) необходимо заменить в ней вектор  $f$  на  $\mu$ .

Для правильного выбора весовых коэффициентов необходимо провести анализ результата. Пусть  $\rho_i = r = \text{const}$ ,  $i = 0, 1, \dots, n$ . Если известна погрешность  $\varepsilon$  в задании  $f_i$  и она носит случайный характер, то ее нужно сравнить со среднеквадратичным уклонением

$$\delta = \left( \sum_{i=0}^n \frac{(\mu_i - f_i)^2}{n+1} \right)^{1/2}.$$

Если  $\delta_k \ll \varepsilon$ , т. е. погрешность аппроксимации существенно меньше погрешности входных данных, то вклад требования «малости кривизны» в решении (и в условии) задачи занижен — вес  $r$  надо уменьшить. Если погрешность аппроксимации существенно больше погрешности входных данных, то сплайн сильно сглажен — надо увеличивать  $r$ . Если  $\delta \approx \varepsilon$ , то значение  $r = r_0$  близко к оптимальному. При любых методах приближения функций следует проявлять осторожность в выборе степени сглаживания, чтобы получился пригодный для дальнейшего использования результат. В зависимости от конкретной цели, после анализа нескольких различных результатов можно отдать предпочтение той или иной степени сглаживания.

Весá  $\rho_i$  можно брать переменными, увеличивая вклад узлов с меньшей «индивидуальной» погрешностью  $\varepsilon_i$  данных  $f_i$ , но соблюдая соотношение  $\delta \approx \varepsilon$ . Например, при найденном  $r = r_0$  в функционале  $\Psi(S)$  весá можно взять в виде  $\rho_i = r_0 r_i$ ,  $r_i \geq 0$ ,  $\sum_{i=0}^n r_i = 1$  и при подсчете уклонения сплайна использовать *взвешенное среднеквадратичное*

$$\delta = \left( \sum_{i=0}^n r_i (\mu_i - f_i)^2 \right)^{1/2}.$$

## Упражнения

1. Таблицы  $\exp x$ ,  $\sin x$ ,  $x^3$  даны от 0 до 1 с шагом 0,01. Каковы наибольшие погрешности линейной и квадратичной интерполяции?

2. Показать, что разделенная разность  $n$ -го порядка многочлена  $n$ -й степени равна коэффициенту при  $x^n$  независимо от выбора узлов  $x_0, x_1, \dots, x_n$ .

3. Для функции  $f(x) = \sqrt{x_0^2 - x^2}$ ,  $x_0 \in [0, 1]$  построить интерполяционный многочлен Лагранжа  $L_2(x)$  по значениям  $f(x)$  в узлах  $x_0 - h$ ,  $x_0$ ,  $x_0 + h$ . Пронаблюдать сходимость значения  $L_2(x_0)$  на последовательности сеток  $h = 0,1, 0,05, 0,025$ . Объяснить результат.

4. Построить интерполяционные многочлены Лагранжа и Эрмита с кратностью всех узлов 2 по значениям одной из функций  $\exp x$ ,  $\sin x$ ,  $x^7$ , взятым в узлах  $x_0 - h$ ,  $x_0 + h$ ,  $x_0 + 3h$  из интервала  $[0, 1]$ . Пронаблюдать сходимость значений многочленов в точке  $x_0$  на последовательности сеток  $h = 0,1, 0,05, 0,025$ .

5. Методом наименьших квадратов найти функцию  $\Phi(x) = \sum_{j=0}^k c_j \varphi_j(x) dx$ , приближающую функцию  $f(x)$ , значения которой заданы со случайной погрешностью. В качестве входных данных взять на интервале  $[0, 1]$  значения  $f(x_i) + (-1)^i \varepsilon \cdot \text{random}(i)$ ,  $i = 0, 1, \dots, n$ , где  $f(x)$  — известная гладкая функция ( $\sin(\pi x/2)$ ,  $\exp(x)$  и т. п.)  $\text{random}(i)$  — датчик случайных чисел со значениями из интервала  $[0, 1]$ . Для различных  $n$  (10, 30, 100) и  $\varepsilon$  (0,01, 0,1, 0,3) подобрать оптимальное  $k$  из условия наилучшего среднеквадратичного приближения. Испытать при большом  $n$  различные наборы  $\varphi_j(x)$ , пронаблюдав соотношение между модулями диагональных и внедиагональных элементов матрицы системы (2.17) в случаях, когда наборы взяты из базисов ортогональных и неортогональных многочленов. В случае использования ортогональных многочленов сделать необходимое линейное преобразование для перевода данных из заданного отрезка в отрезок  $[-1, 1]$ . Сравнить во всех случаях построенную аппроксимирующую функцию  $\Phi(x)$  и  $f(x)$  на подробной равномерной сетке  $x_i$ ,  $i = 0, 1, \dots, n$  вычислив максимум отклонения  $\delta = \max_i \|\Phi(x_i) - f(x_i)\|$ .

6. Построить интерполирующий сплайн по значениям одной из функций  $\exp x$ ,  $\sin x$ ,  $x^7$ , взятым в узлах  $x_i$ ,

$i = 0, 1, \dots, n$  на интервале  $[0, 1]$  на равномерной и неравномерной сетках; на более мелкой сетке в промежутке между узлами интерполяции вычислить значения сплайна, первой и второй производных от сплайна; оценить погрешность результата интерполирования сравнением со значениями  $f(x)$  и ее производных. Пронаблюдать сходимость результатов на последовательности сеток  $h = 0,1, 0,05, 0,025$ .

7. Построить сглаживающий сплайн; в качестве входных данных взять на интервале  $[0, 1]$  значения  $f(x_i) + (-1)^i \varepsilon \cdot \text{random}(i)$ ,  $i = 0, 1, \dots, n$ , где  $f(x)$  — известная гладкая функция ( $\sin(\pi x)$ ,  $\exp x$  и т. п.),  $\text{random}(i)$  — датчик случайных чисел со значениями из интервала  $[0, 1]$ . Для различных  $n$  (10, 30, 100) и  $\varepsilon$  (0,01, 0,1, 0,3) подобрать оптимальное  $r_0$  (обозначения и способ решения задачи см. в п. 2.4.3) из условия  $\delta \approx \varepsilon$  путем построения нескольких сплайнов с различными весами  $r$ . Пронаблюдать степень сглаживания входных данных и уклонение сплайна, взяв в качестве расстояния между  $S(x)$  и  $f(x)$  величину  $\left( \int_{x_0}^{x_n} (S(x) - f(x))^2 dx \right)^{1/2}$ , которую следует вычислить с необходимой точностью по квадратурной формуле. При небольшом  $n$ , зная величину погрешности в каждом узле, построить несколько сплайнов, придавая бóльший вес узлам с бóльшими погрешностями; выбрать из построенных решений наименее уклоняющееся.



## ЧИСЛЕННОЕ ИНТЕГРИРОВАНИЕ

Известно, что вычисление определенного интеграла  $I = \int_a^b f(x)dx$  через первообразную подынтегральной функции с помощью формулы Ньютона — Лейбница зачастую сопряжено с немалыми трудностями. Более того, первообразные некоторых элементарных функций сами таковыми не являются. Наконец, функция  $f$  может быть задана не аналитически, а таблично, т. е. на конечном множестве точек. В силу этих причин приходится прибегать к приближенному вычислению интеграла посредством какой-либо *квадратурной формулы*, то есть приближенного равенства

$$\int_a^b f(x)dx \approx \sum_{k=0}^n c_k f(x_k), \quad (3.1)$$

где  $c_k$  — постоянные, называемые *коэффициентами квадратурной формулы*, а  $x_k$  — точки отрезка  $[a, b]$ , называемые *узлами квадратурной формулы*. Совокупность узлов называется *сеткой*.

Разность

$$\Psi = \int_a^b f(x)dx - \sum_{k=0}^n c_k f(x_k)$$

называется *погрешностью квадратурной формулы*. Как правило, оценка погрешности имеет вид

$$|\Psi| \leq C(b-a)^k, \quad (3.2)$$

где  $k > 1$ ,  $C$  — число, зависящее от типа квадратурной формулы и от подынтегральной функции  $f$ , но не зависящее от длины отрезка  $[a, b]$ . При оценке погрешности функция  $f(x)$  предполагается достаточно гладкой.

В теории квадратурных формул используется свойство аддитивности определенного интеграла: для любой точки  $d \in [a, b]$  выполняется равенство

$$\int_a^b f(x) dx = \int_a^d f(x) dx + \int_d^b f(x) dx.$$

Пусть для погрешности квадратурной формулы (3.1) имеет место оценка вида (3.2). Разобьем отрезок  $[a, b]$  точками  $d_1, d_2, \dots, d_{N-1}$  на  $N$  частичных отрезков длины  $h = (b - a)/N$ :  $[d_0, d_1], [d_1, d_2], \dots, [d_{N-1}, d_N]$  (полагаем  $d_0 = a, d_N = b$ ) и вычислим интеграл на каждом из них по формуле (3.1), при этом погрешность на любом частичном отрезке оценивается следующим образом:

$$|\psi_N| \leq C \left( \frac{b - a}{N} \right)^k = Ch^k.$$

Используя свойство аддитивности интеграла, оценим погрешность интегрирования на всем отрезке  $[a, b]$ :

$$|\Psi_N| \leq C N \left( \frac{b - a}{N} \right)^k = C \frac{(b - a)^k}{N^{k-1}} = C(b - a)h^{k-1}. \quad (3.3)$$

Число  $k - 1$  называют *порядком точности квадратурной формулы*.

Итак, разбиение отрезка интегрирования на  $N$  равных частей привело к уменьшению погрешности в  $N^{k-1}$  раз. Взяв  $N$  достаточно большим, мы можем сделать погрешность интегрирования сколь угодно малой.

### § 3.1. КВАДРАТУРНЫЕ ФОРМУЛЫ ИНТЕРПОЛЯЦИОННОГО ТИПА

Простейшим приемом получения квадратурных формул является замена подынтегральной функции  $f(x)$  на отрезке  $[a, b]$  интерполяционным многочленом, построенным по узлам квадратурной формулы вида (3.1), с последующим интегрированием этого многочлена. Полученные таким способом формулы называются *квадратурными формулами интерполяционного типа*. Если при этом сетка равномерная, то мы имеем дело с *формулами Ньютона — Котеса*. Далее будем полагать  $x_0 = a, x_n = b, x_k - x_{k-1} = (b - a)/n, k = 1, 2, \dots, n$  (исключение составляет случай  $n = 0$ , когда мы считаем  $x_0 = (a + b)/2$ ).

Ниже будет осуществлен вывод формул интерполяционного типа для  $n = 0, 1, 2$ .

### 3.1.1. ФОРМУЛА ПРЯМОУГОЛЬНИКОВ

При  $n = 0$  полагаем  $x_0 = (a + b)/2$ , то есть  $f(x) \approx f(\frac{a+b}{2})$ . Тогда искомая квадратурная формула имеет вид

$$\int_a^b f(x)dx \approx (b - a)f(\frac{a+b}{2}). \quad (3.4)$$

Она называется *формулой прямоугольников*, ибо приближенное вычисление интеграла фактически сводится к нахождению площади прямоугольника с длинами сторон  $(b - a)$  и  $f(\frac{a+b}{2})$ .

Оценим погрешность формулы прямоугольников, используя разложение функции  $f(x)$  в ряд Тейлора в окрестности точки  $(a + b)/2$ :

$$f(x) = f(\frac{a+b}{2}) + f'(\frac{a+b}{2})(x - \frac{a+b}{2}) + f''(\xi) \frac{(x - \frac{a+b}{2})^2}{2},$$

где  $\xi \in [a, b]$ . С учетом очевидного равенства

$$\int_a^b (x - \frac{a+b}{2})^{2k-1} dx = 0, \quad k = 1, 2, \dots, \quad (3.5)$$

получим, что

$$\begin{aligned} \Psi &= \int_a^b f(x)dx - (b - a)f(\frac{a+b}{2}) = \\ &= \int_a^b (f(x) - f(\frac{a+b}{2}))dx = \int_a^b f''(\xi) \frac{(x - \frac{a+b}{2})^2}{2} dx. \end{aligned}$$

Вводя обозначение

$$M_k = \max_{x \in [a, b]} |f^{(k)}(x)|, \quad (3.6)$$

выпишем оценку погрешности:

$$|\Psi| \leq M_2 \int_a^b \frac{(x - \frac{a+b}{2})^2}{2} dx = M_2 \frac{(b-a)^3}{24}. \quad (3.7)$$

Отсюда следует, что для многочленов степени не выше 1 формула прямоугольников дает точное равенство.

Оценка (3.7) является неулучшаемой, ибо для функции  $f(x) = (x - \frac{a+b}{2})^2$  имеем  $M_2 = 2$ ,  $f(\frac{a+b}{2}) = 0$  и

$$\int_a^b f(x) dx - (b-a)f(\frac{a+b}{2}) = \frac{(b-a)^3}{12} = M_2 \frac{(b-a)^3}{24},$$

то есть (3.7) выполняется со знаком равенства.

При разбиении отрезка  $[a, b]$  на  $N$  равных частей длины  $h = (b-a)/N$ , как это было описано выше, применение на каждом частичном отрезке формулы вида (3.4) даст *составную формулу прямоугольников*:

$$\int_a^b f(x) dx \approx h \sum_{i=1}^N f(d_{i-1/2}),$$

где  $d_{i-1/2} = d_i - h/2$ . Ее погрешность, согласно (3.3), оценивается следующим образом:

$$|\Psi_N| \leq M_2 \frac{(b-a)h^2}{24}.$$

Тем самым формула прямоугольников имеет второй порядок точности.

**Задача 3.1.** Доказать, что при  $x_0 = a$  или  $x_0 = b$  формула прямоугольников имеет лишь первый порядок точности.

### 3.1.2. ФОРМУЛА ТРАПЕЦИЙ

Заменяя подынтегральную функцию интерполяционным многочленом первой степени, построенным по узлам  $x_0 = a$ ,  $x_1 = b$ , который имеет вид

$$L_1(x) = \frac{1}{b-a} \left( (x-a)f(b) - (x-b)f(a) \right),$$

получим *формулу трапеций*:

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

Для оценки ее погрешности воспользуемся формулой (2.2), согласно которой

$$f(x) - L_1(x) = f''(\xi(x)) \frac{(x-a)(x-b)}{2}.$$

Отсюда следует, что

$$\Psi = \int_a^b f''(\xi(x)) \frac{(x-a)(x-b)}{2} dx,$$

то есть

$$|\Psi| \leq M_2 \frac{(b-a)^3}{12}.$$

Эта оценка неулучшаема, так как в ней достигается равенство, например, при  $f(x) = (x-a)^2$ .

Выпишем *составную формулу трапеций*:

$$\int_a^b f(x)dx \approx h \left( \frac{1}{2}f(d_0) + \sum_{i=1}^{N-1} f(d_i) + \frac{1}{2}f(d_N) \right),$$

где  $h = (b-a)/N$ , и оценим ее погрешность:

$$|\Psi_N| \leq M_2 \frac{(b-a)h^2}{12}.$$

Итак, формула трапеций, как и формула прямоугольников, имеет второй порядок точности, но ее погрешность оценивается величиной в два раза большей.

### 3.1.3. ФОРМУЛА СИМПСОНА

Получим квадратурную формулу интерполяционного типа с тремя узлами  $x_0 = a$ ,  $x_1 = (a+b)/2$ ,  $x_2 = b$ , заменив функцию  $f(x)$  многочленом Лагранжа второй степени, имеющим вид

$$L_2(x) = \frac{2}{(b-a)^2} \left( \left( x - \frac{a+b}{2} \right) (x-b) f(a) - \right.$$

$$-2(x-a)(x-b)f\left(\frac{a+b}{2}\right) + (x-a)\left(x-\frac{a+b}{2}\right)f(b)\Big).$$

Интегрируя это выражение, получим приближенное равенство

$$\int_a^b f(x)dx \approx \frac{b-a}{6} \left( f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right),$$

называемое *формулой Симпсона* или *формулой парабол*.

**Задача 3.2.** Доказать, что формула Симпсона точна для любого многочлена третьей степени.

Для оценки погрешности формулы Симпсона рассмотрим интерполяционный многочлен Эрмита третьей степени  $H_3(x)$  такой, что

$$H_3(a) = f(a), \quad H_3\left(\frac{a+b}{2}\right) = f\left(\frac{a+b}{2}\right),$$

$$H'_3\left(\frac{a+b}{2}\right) = f'\left(\frac{a+b}{2}\right), \quad H_3(b) = f(b).$$

В п. 2.1.4 доказаны существование и единственность такого многочлена, а также получена явная форма его остатка. В данном конкретном случае

$$f(x) - H_3(x) = \frac{f^{IV}(\xi)}{24}(x-a)\left(x-\frac{a+b}{2}\right)^2(x-b).$$

Используя результат задачи 3.2, получим равенство

$$\begin{aligned} \int_a^b H_3(x)dx &= \frac{b-a}{6} \left( H_3(a) + 4H_3\left(\frac{a+b}{2}\right) + H_3(b) \right) = \\ &= \frac{b-a}{6} \left( f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right). \end{aligned}$$

Отсюда

$$\begin{aligned} \Psi &= \int_a^b f(x)dx - \frac{b-a}{6} \left( f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right) = \\ &= \int_a^b (f(x) - H_3(x))dx. \end{aligned}$$

Таким образом,

$$|\Psi| \leq \frac{M_4}{24} \left| \int_a^b (x-a) \left(x - \frac{a+b}{2}\right)^2 (x-b) dx \right| = \frac{M_4}{24} \frac{(b-a)^5}{120}$$

и, окончательно,

$$|\Psi| \leq M_4 \frac{(b-a)^5}{2880}.$$

Разбиение отрезка на  $N$  равных частей длины  $h = (b-a)/N$  даст *составную формулу Симпсона*:

$$\int_a^b f(x) dx \approx \frac{h}{6} \sum_{i=1}^N (f(d_{i-1}) + 4f(d_{i-1/2}) + f(d_i)),$$

где  $d_{i-1/2} = d_i - h/2$ . Оценим ее погрешность:

$$|\Psi_N| \leq M_4 \frac{(b-a)h^4}{2880}.$$

Таким образом, формула Симпсона, имеющая четвертый порядок точности, существенно точнее, чем формулы прямоугольников и трапеций.

### 3.1.4. НЕКОТОРЫЕ ОБОБЩЕНИЯ

Как уже говорилось, квадратурная формула интерполяционного типа получается заменой подынтегральной функции  $f(x)$  интерполяционным многочленом Лагранжа,

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

где

$$\omega_n(x) = \prod_{j=0}^n (x-x_j), \quad \omega'_n(x_k) = \prod_{j \neq k}^n (x_k-x_j)$$

(отметим, что в этом пункте мы не требуем равномерности сетки). Тогда коэффициенты квадратурной формулы (3.1) будут иметь вид

$$c_k = \int_a^b \frac{\omega_n(x)}{(x-x_k) \omega'_n(x_k)} dx, \quad k = 0, 1, \dots, n, \quad (3.8)$$

а ее погрешность равна

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

где  $r_n(x)$  — погрешность интерполирования. В п. 2.1.3 показано, что

$$r_n(x) = \frac{f^{(n+1)}(\xi(x))\omega_n(x)}{(n+1)!},$$

откуда

$$\Psi_n = \frac{1}{(n+1)!} \int_a^b f^{(n+1)}(\xi(x))\omega_n(x) dx.$$

В итоге имеем оценку погрешности квадратурной формулы интерполяционного типа:

$$|\Psi_n| \leq \frac{M_{n+1}}{(n+1)!} \int_a^b |\omega_n(x)| dx. \quad (3.9)$$

Отсюда видно, что квадратурная формула интерполяционного типа, построенная по  $n+1$  узлу, точна для любого многочлена степени  $n$ .

Наоборот, если квадратурная формула вида (3.1) точна для любого многочлена степени  $n$ , то она является квадратурной формулой интерполяционного типа, т. е. ее коэффициенты  $c_k$  вычисляются согласно (3.8). Действительно, рассматривая базисные многочлены Лагранжа

$$\varphi_k(x) = \frac{\omega_n(x)}{(x - x_k)\omega'_n(x_k)},$$

имеющие степень  $n$ , и используя их свойства, описанные в п. 2.1.1, получим точные равенства

$$\int_a^b \varphi_k(x) dx = \sum_{l=0}^n c_l \varphi_k(x_l) = \sum_{l=0}^n c_l \delta_{kl} = c_k,$$

что и требовалось доказать.

### 3.1.5. СИММЕТРИЧНЫЕ ФОРМУЛЫ

Легко убедиться, что оценки погрешности для формул прямоугольников и Симпсона, полученные нами в п. 3.1.1 и



п. 3.1.3, более точны, чем оценки для этих же формул, следующие из (3.9). Так, согласно результату задачи 3.2, формула Симпсона точна на любом многочлене третьей степени, в частности на функции  $f(x) = x^3$ . С другой стороны, погрешность  $\Psi_2$  формулы Симпсона для интеграла  $\int_{-1}^1 x^3 dx$ , вычисленная по формуле (3.9), не равна 0.

**Задача 3.3.** Доказать, что  $|\Psi_2| = 1/2$ .

Оказывается, причина повышенной точности формул прямоугольников и Симпсона заключается в их симметричности. Вообще, формула вида (3.1) называется *симметричной*, если выполняется совокупность условий:

- 1)  $n$  четно;
- 2) узлы квадратурной формулы расположены симметрично относительно середины отрезка  $[a, b]$ , то есть

$$\frac{a+b}{2} - x_k = x_{n-k} - \frac{a+b}{2}, \quad k = 0, 1, \dots, n/2, \quad (3.10)$$

$$3) \quad c_k = c_{n-k}, \quad k = 0, 1, \dots, n/2. \quad (3.11)$$

Покажем, что для квадратурных формул интерполяционного типа при четном  $n$  условие (3.11) является следствием условия (3.10). Пусть  $x_{n/2} = (a+b)/2$ . Тогда, учитывая (3.10), имеем

$$\begin{aligned} \omega_n(x) &= (x - x_{n/2}) \prod_{k=0}^{n/2-1} (x - x_k)(x - x_{n-k}) = \\ &= (x - x_{n/2}) \prod_{k=0}^{n/2-1} [(x - x_{n/2})^2 - (x_k - x_{n/2})^2], \end{aligned}$$

откуда при любом  $t$

$$\omega_n(x_{n/2} + t) = t \prod_{k=0}^{n/2-1} [t^2 - (x_k - x_{n/2})^2] = -\omega_n(x_{n/2} - t),$$

то есть функция  $\omega_n(x)$  нечетна относительно точки  $x_{n/2}$ , и, следовательно, ее производная  $\omega'_n(x)$  — четная функция относительно  $x_{n/2}$ . В частности,

$$\omega'_n(x_{n-k}) = \omega'_n(x_k).$$

Так как коэффициенты  $c_k$  подсчитываются по формуле (3.8), получим

$$\begin{aligned} c_k - c_{n-k} &= \int_a^b \left( \frac{\omega_n(x)}{(x - x_k) \omega'_n(x_k)} - \frac{\omega_n(x)}{(x - x_{n-k}) \omega'_n(x_{n-k})} \right) dx = \\ &= \int_a^b \frac{\omega_n(x)}{\omega'_n(x_k)} \frac{x_k - x_{n-k}}{(x - x_k)(x - x_{n-k})} dx = \\ &= \int_a^b \frac{\omega_n(x)}{\omega'_n(x_k)} \frac{x_k - x_{n-k}}{[(x - x_{n/2})^2 - (x_k - x_{n/2})^2]} dx. \end{aligned}$$

Таким образом, под интегралом стоит функция, нечетная относительно точки  $x_{n/2}$  — середины отрезка  $[a, b]$ , то есть интеграл равен нулю. Значит, равенство (3.11) выполняется.

Теперь мы можем показать, что наличие симметрии повышает точность квадратурных формул.

**Теорема 3.1.** Пусть расположение узлов квадратурной формулы интерполяционного типа (3.1) удовлетворяет условию (3.10), где  $n$  — четное число, и формула точна для любого многочлена степени  $n$ . Тогда она точна и для любого многочлена степени  $n + 1$ .

**Доказательство.** Достаточно показать, что формула точна хотя бы для одного многочлена степени  $n + 1$ , например для  $f(x) = (x - x_{n/2})^{n+1}$ .

Так как подынтегральная функция нечетна относительно точки  $x_{n/2}$ , то

$$\int_a^b (x - x_{n/2})^{n+1} dx = 0,$$

поэтому нам требуется доказать, что

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

Очевидно,

$$\sum_{k=0}^n c_k f(x_k) = \sum_{k=0}^{n/2-1} c_k (x_k - x_{n/2})^{n+1} + \sum_{k=n/2+1}^n c_k (x_k - x_{n/2})^{n+1}. \quad (3.12)$$

Преобразуем вторую сумму, используя условие (3.10) и четность числа  $n$ :

$$\begin{aligned} \sum_{k=n/2+1}^n c_k (x_k - x_{n/2})^{n+1} &= \sum_{k=n/2+1}^n c_k (x_{n/2} - x_{n-k})^{n+1} = \\ &= \sum_{l=0}^{n/2-1} c_{n-l} (x_{n/2} - x_l)^{n+1} = - \sum_{k=0}^{n/2-1} c_{n-k} (x_k - x_{n/2})^{n+1}. \end{aligned}$$

Подставляя последнее выражение в (3.12) и используя равенство (3.11), получим, что рассматриваемая квадратурная сумма равна нулю.

### 3.1.6. АПОСТЕРИОРНАЯ ОЦЕНКА ПОГРЕШНОСТИ МЕТОДОМ РУНГЕ

Практическое применение полученных нами априорных формул оценки погрешности, как правило, затруднено тем, что в них входит величина

$$M_k = \max_{x \in [a, b]} |f^{(k)}(x)|,$$

которая, вообще говоря, заранее неизвестна. Для того чтобы правильно выбрать величину шага сетки  $h$ , позволяющую достичь заданную точность, нужно уметь оценивать погрешность апостериорно, то есть после проведения расчета.

Одним из приемов апостериорной оценки погрешности является *метод Рунге*, который мы рассмотрим сначала на примере формулы трапеций. Обозначим для краткости длину частичного отрезка  $[d_{i-1}, d_i]$  через  $h_i$ . Тогда, согласно формуле трапеций,

$$I_i = \int_{d_{i-1}}^{d_i} f(x) dx \approx \frac{f(d_{i-1}) + f(d_i)}{2} h_i = I_{h,i}.$$

Из априорной оценки погрешности для достаточно гладких функций следует, что

$$I_i - I_{h,i} \approx c_i h_i^3, \quad (3.13)$$

причем константа  $c_i$ , не зависящая от  $h$ , заранее неизвестна. Разбив отрезок  $[d_{i-1}, d_i]$  на две равные части и проводя расчет на каждой из них с шагом  $h_i/2$ , получим

$$I_{h/2,i} = \left( f(d_{i-1}) + 2f\left(\frac{d_{i-1}+d_i}{2}\right) + f(d_i) \right) \frac{h_i}{4}.$$

Аналогично (3.13) имеем

$$I_i - I_{h/2,i} \approx 2c_i \left( \frac{h_i}{2} \right)^3. \quad (3.14)$$

Исключив из соотношений (3.13) и (3.14) константу  $c_i$ , можно получить оценку погрешности, содержащую лишь известные величины  $I_{h,i}$  и  $I_{h/2,i}$ :

$$I_i - I_{h,i} \approx \frac{4}{3}(I_{h/2,i} - I_{h,i}),$$

$$I_i - I_{h/2,i} = \frac{1}{4}(I_i - I_{h,i}) \approx \frac{1}{3}(I_{h/2,i} - I_{h,i}).$$

Отсюда следует, что путем уточнения формулы трапеций с помощью правила Рунге легко вывести формулу Симпсона. Действительно, рассматривая отрезок  $[d_0, d_1]$ , имеем

$$\begin{aligned} I_i \approx \frac{1}{3}(4I_{h/2,i} - I_{h,i}) &= \frac{1}{3} \left[ 4 \frac{h_i}{2} \left( \frac{1}{2}f(d_0) + f(d_{1/2}) + \frac{1}{2}f(d_1) \right) - \right. \\ &\left. - h_i \left( \frac{1}{2}f(d_0) + \frac{1}{2}f(d_1) \right) \right] = \frac{h_i}{6}(f(d_0) + 4f(d_{1/2}) + f(d_1)). \end{aligned}$$

В общем случае, если  $I_i - I_{h,i} \approx c_i h_i^m$ , то

$$I_i - I_{h/2,i} \approx 2c_i \left( \frac{h_i}{2} \right)^m,$$

откуда

$$I_i - I_{h,i} \approx \frac{2^{m-1}(I_{h/2,i} - I_{h,i})}{2^{m-1} - 1},$$

$$I_i - I_{h/2,i} \approx \frac{I_{h/2,i} - I_{h,i}}{2^{m-1} - 1}.$$

Опишем, как с помощью метода Рунге можно вычислить интеграл  $I$  на отрезке  $[a, b]$  с наперед заданной точностью  $\varepsilon$ . Разобьем отрезок  $[a, b]$  на  $N$  отрезков  $[d_{i-1}, d_i]$  длины  $h_i$ ,  $i = 1, \dots, N$ , и применим на каждом из них одну и ту же квадратурную формулу. Получим

$$I \approx I_h = \sum_{i=1}^N I_{h,i}.$$

Повторив вычисления с шагом  $h_i/2$ , оценим погрешность по правилу Рунге и проверим выполнение неравенств

$$\frac{|I_{h/2,i} - I_{h,i}|}{2^{m-1} - 1} \leq \frac{\varepsilon h_i}{b - a}, \quad i = 1, \dots, N.$$

Если все эти неравенства истинны, то

$$|I - I_{h/2}| \leq \frac{\varepsilon}{b - a} \sum_{i=1}^N h_i = \varepsilon, \quad (3.15)$$

откуда следует, что интеграл подсчитан с заданной точностью. В противном случае на тех из отрезков, где не выполнено неравенство (3.15), нужно уменьшить шаг сетки еще в 2 раза и снова оценить погрешность. Указанная процедура продолжается до выполнения неравенства вида (3.15) для всех отрезков.

## § 3.2. КВАДРАТУРНЫЕ ФОРМУЛЫ ГАУССА

### 3.2.1. ПОСТАНОВКА ЗАДАЧИ

В предыдущем параграфе показано, что квадратурная формула интерполяционного типа с  $n$  узлами точна, вообще говоря, для многочленов степени  $n - 1$ . Можно поставить следующую задачу: построить квадратурную формулу вида

$$\int_a^b f(x) dx \approx \sum_{k=1}^n c_k f(x_k), \quad (3.16)$$

которая при заданном числе узлов  $n$  была бы точна для многочленов наиболее высокой степени. Такие формулы называются *формулами Гаусса*. Отметим, что в отличие от формулы (3.1) нумерация узлов начинается с  $k = 1$ .

Очевидно, требование точного равенства в формуле (3.16) для любого многочлена степени  $m$  эквивалентно тому, что указанная формула точна для одночленов  $f(x) = x^l$ ,  $l = 0, 1, \dots, m$ , то есть

$$\sum_{k=1}^n c_k x_k^l = \int_a^b x^l dx, \quad l = 0, 1, \dots, m. \quad (3.17)$$

Тем самым получена нелинейная система  $m + 1$  уравнений относительно  $2n$  неизвестных  $c_1, c_2, \dots, c_n$ ;  $x_1, x_2, \dots, x_n$ . Число уравнений равняется числу неизвестных при  $m = 2n - 1$ . Именно таково значение наиболее высокой степени многочленов, для которых точна формула (3.16). Для доказательства этого утверждения мы покажем, что система (3.17) при  $m = 2n - 1$  имеет единственное решение, но сначала приведем два частных случая формул Гаусса.

Если  $n = 1$ , то и  $m = 1$ , а система (3.17) примет вид

$$c_1 = b - a, \quad c_1 x_1 = \frac{1}{2}(b^2 - a^2),$$

откуда  $x_1 = \frac{a+b}{2}$ . Таким образом, мы получили формулу прямоугольников, подробно рассмотренную в п. 3.1.1:

$$\int_a^b f(x) dx \approx (b - a) f\left(\frac{a+b}{2}\right)$$

Значит, формула прямоугольников является формулой Гаусса с  $n = 1$ .

Запишем систему (3.17) при  $n = 2$ ,  $m = 3$ :

$$c_1 + c_2 = b - a, \quad c_1 x_1 + c_2 x_2 = \frac{1}{2}(b^2 - a^2),$$

$$c_1 x_1^2 + c_2 x_2^2 = \frac{1}{3}(b^3 - a^3), \quad c_1 x_1^3 + c_2 x_2^3 = \frac{1}{4}(b^4 - a^4).$$

Тогда

$$c_1 = c_2 = \frac{b-a}{2}, \quad x_{1,2} = \frac{a+b}{2} \mp \frac{\sqrt{3}}{3} \frac{b-a}{2}.$$

Эта формула точна для многочленов третьей степени.

Ввиду высокой точности и экономичности метод Гаусса часто применяется на практике и включен во многие библиотеки программ. Его нередко используют в виде прямого произведения одномерных квадратур для вычисления многомерных интегралов.

### 3.2.2. СУЩЕСТВОВАНИЕ И ЕДИНСТВЕННОСТЬ ФОРМУЛ ГАУССА

Введем многочлен

$$\omega(x) = (x - x_1)(x - x_2) \dots (x - x_n),$$

где  $x_1, x_2, \dots, x_n$  — узлы квадратурной формулы (3.16).

**Теорема 3.2.** *Квадратурная формула (3.16) точна для любого многочлена степени  $m = 2n - 1$  тогда и только тогда, когда выполнены два условия:*

1) *многочлен  $\omega(x)$  ортогонален любому многочлену  $q(x)$  степени меньше  $n$ , то есть*

$$\int_a^b \omega(x)q(x)dx = 0;$$

2) *формула (3.16) является формулой интерполяционного типа, то есть*

$$c_k = \int_a^b \frac{\omega(x)}{(x - x_k) \omega'(x_k)} dx, \quad k = 1, 2, \dots, n.$$

**Доказательство.** **Необходимость.** Если формула (3.16) точна для любого многочлена степени  $m = 2n - 1$ , то она точна, в частности, и для многочлена  $\omega(x)q(x)$ , степень которого не выше  $2n - 1$ , то есть

$$\int_a^b \omega(x)q(x)dx = \sum_{k=1}^n c_k \omega(x_k)q(x_k) = 0.$$

Требование 2) выполняется в силу утверждения из п. 3.1.4 (если квадратурная формула вида (3.16) точна для любого

многочлена степени  $n - 1$ , то она является квадратурной формулой интерполяционного типа).

**Д о с т а т о ч н о с т ь.** Согласно теореме о делимости многочленов, произвольный многочлен  $f(x)$  степени  $2n - 1$  может быть представлен в виде

$$f(x) = \omega(x)q(x) + r(x),$$

где  $q(x)$ ,  $r(x)$  — многочлены степени не выше  $n - 1$ . В силу предположения 2)

$$\int_a^b f(x)dx = \int_a^b \omega(x)q(x)dx + \int_a^b r(x)dx = \int_a^b r(x)dx.$$

Так как формула (3.16) является квадратурной формулой интерполяционного типа, построенной по  $n$  узлам, то, согласно п. 3.1.4, она точна на многочленах степени не выше  $n - 1$ , в частности на  $r(x)$ . Поэтому

$$\begin{aligned} \int_a^b r(x)dx &= \sum_{k=1}^n c_k r(x_k) = \\ &= \sum_{k=1}^n c_k (f(x_k) - \omega(x_k)q(x_k)) = \sum_{k=1}^n c_k f(x_k). \end{aligned}$$

Значит,

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

то есть формула (3.16) точна для любого многочлена степени  $2n - 1$ .

Очевидно, условие 1) эквивалентно равенствам

$$\int_a^b \omega(x)x^\alpha dx = 0, \quad \alpha = 0, 1, \dots, n - 1, \quad (3.18)$$

представляющим собой систему  $n$  нелинейных уравнений относительно неизвестных  $x_1, x_2, \dots, x_n$ . Решив ее, мы найдем узлы интерполяции, после чего коэффициенты  $c_k$  вычисляются по формуле из условия 2).



Для доказательства существования и единственности решения системы (3.18) достаточно доказать существование и единственность многочлена  $\omega(x)$ , ортогонального любому многочлену степени меньше  $n$ , показав при этом, что все его корни различны и лежат на отрезке  $[a, b]$ .

Запишем искомым многочлен  $\omega(x)$  в виде

$$\omega(x) = a_0 + a_1x + \dots + a_{n-1}x^{n-1} + x^n$$

и воспользуемся условием ортогональности (3.18):

$$\int_a^b (a_0 + a_1x + \dots + a_{n-1}x^{n-1} + x^n) x^\alpha dx = 0, \quad \alpha = 0, 1, \dots, n-1.$$

Мы получили неоднородную систему  $n$  линейных уравнений относительно  $n$  неизвестных  $a_1, a_2, \dots, a_{n-1}$ . Чтобы установить ее однозначную разрешимость, достаточно показать, что однородная система

$$\int_a^b (a_0 + a_1x + \dots + a_{n-1}x^{n-1}) x^\alpha dx = 0, \quad \alpha = 0, 1, \dots, n-1, \quad (3.19)$$

имеет только нулевое решение  $a_0 = a_1 = \dots = a_{n-1} = 0$ . Умножая уравнение системы (3.19), имеющее номер  $\alpha$ , на  $a_\alpha$  и производя суммирование по всем  $\alpha$ , получим

$$\sum_{\alpha=0}^{n-1} a_\alpha \int_a^b \left( \sum_{k=0}^{n-1} a_k x^k \right) x^\alpha dx = \int_a^b \left( \sum_{\alpha=0}^{n-1} \sum_{k=0}^{n-1} a_\alpha a_k x^k x^\alpha \right) dx = 0.$$

Отсюда следует, что

$$\int_a^b \left( \sum_{l=0}^{n-1} a_l x^l \right)^2 dx = 0,$$

то есть  $a_0 = a_1 = \dots = a_{n-1} = 0$ .

Таким образом, мы показали, что существует единственный многочлен  $\omega(x)$  степени  $n$  со старшим коэффициентом 1, ортогональный любому многочлену степени  $n-1$ . Осталось доказать, что все его корни различны и лежат на отрезке  $[a, b]$ .

Предположим, что среди корней многочлена  $\omega(x)$ , лежащих на  $[a, b]$ , имеется  $m$  различных корней  $\xi_1, \xi_2, \dots, \xi_m$  нечетной кратности  $\alpha_1, \alpha_2, \dots, \alpha_m$  (очевидно,  $m \leq n$ ). Требуется доказать, что  $m = n$ . Запишем  $\omega(x)$  в виде

$$\omega(x) = (x - \xi_1)^{\alpha_1} (x - \xi_2)^{\alpha_2} \dots (x - \xi_m)^{\alpha_m} r(x),$$

где  $r(x)$  — многочлен, не меняющий знак на  $[a, b]$ . Рассмотрим интеграл

$$\begin{aligned} I &= \int_a^b \omega(x)(x - \xi_1) \dots (x - \xi_m) dx = \\ &= \int_a^b (x - \xi_1)^{\alpha_1+1} \dots (x - \xi_m)^{\alpha_m+1} r(x) dx. \end{aligned}$$

Так как  $\alpha_1 + 1, \dots, \alpha_m + 1$  — четные числа и функция  $r(x)$  не меняет знак на  $[a, b]$ , то  $I \neq 0$ . С другой стороны, если  $m < n$ , то  $q(x) = (x - \xi_1)(x - \xi_2) \dots (x - \xi_m)$  — многочлен степени меньше  $n$ . Тогда  $\omega(x)$  ортогонален  $q(x)$ , то есть  $I = 0$ . Значит,  $m = n$ , что и требовалось установить.

Таким образом, мы доказали, что для любого  $n$  существует единственная квадратурная формула, точная для любого многочлена степени  $2n - 1$ .

### 3.2.3. СВОЙСТВА КВАДРАТУРНЫХ ФОРМУЛ ГАУССА

Заметим, что  $2n - 1$  — наивысшая степень точности формул Гаусса. Действительно, для многочлена  $\omega^2(x)$ , имеющего степень  $2n$ , выполняются равенства

$$\int_a^b \omega^2(x) dx > 0$$

и

$$\sum_{k=0}^n c_k \omega^2(x_k) = 0,$$

то есть для этого многочлена формула Гаусса с  $n$  узлами не является точной.

Найдем погрешность формул Гаусса в случае, когда подынтегральная функция  $f(x)$  достаточно гладкая. Для этого рассмотрим интерполяционный многочлен Эрмита  $H(x)$  с узлами  $x_1, x_2, \dots, x_n$  кратности 2:

$$H(x_k) = f(x_k), \quad H'(x_k) = f'(x_k), \quad k = 1, 2, \dots, n.$$

Свойства этого многочлена описаны в п. 2.1.4. В частности, из формулы (2.6) следует, что

$$f(x) - H(x) = \frac{f^{(2n)}(\xi(x))}{(2n)!} \prod_{i=1}^k (x - x_i)^2 = \frac{f^{(2n)}(\xi(x))}{(2n)!} \omega^2(x).$$

Многочлен  $H(x)$  имеет степень  $2n - 1$ . Следовательно,

$$\begin{aligned} \int_a^b f(x) dx &= \int_a^b H(x) dx + \int_a^b \frac{f^{(2n)}(\xi(x))}{(2n)!} \omega^2(x) dx = \\ &= \sum_{k=0}^n c_k H(x_k) + \int_a^b \frac{f^{(2n)}(\xi(x))}{(2n)!} \omega^2(x) dx = \\ &= \sum_{k=0}^n c_k f(x_k) + \int_a^b \frac{f^{(2n)}(\xi(x))}{(2n)!} \omega^2(x) dx. \end{aligned}$$

Таким образом, погрешность имеет вид

$$\Psi(f) = \int_a^b \frac{f^{(2n)}(\xi(x))}{(2n)!} \omega^2(x) dx.$$

Так как функция  $\omega^2(x)$  знакопостоянна, то, согласно обобщенной теореме о среднем, существует  $\eta \in (a, b)$  такое, что

$$\Psi(f) = \frac{f^{(2n)}(\eta)}{(2n)!} \int_a^b \omega^2(x) dx.$$

Отсюда следует, что формулы Гаусса с большим числом узлов целесообразно применять лишь для достаточно гладких функций.

**Задача 3.4.** Доказать, что

$$\Psi(f) = f^{(2n)}(\eta) \frac{(b-a)^{2n+1} (n!)^4}{[(2n)!]^3 (2n+1)}.$$

**Указание.** Показать, что

$$\omega(x) = \frac{n!}{(2n)!} \frac{d^n}{dx^n} [(x-a)^n (x-b)^n],$$

и  $2n$  раз воспользоваться формулой интегрирования по частям.

**Задача 3.5.** Оценить погрешность формул Гаусса при  $n = 1$  и  $n = 2$ .

Чтобы не пересчитывать каждый раз заново коэффициенты и узлы интегрирования, существуют *таблицы значений* для различных  $n$  на отрезке  $[-1, 1]$ . Произвольный отрезок  $[a, b]$  может быть отображен на этот отрезок простой заменой переменной интегрирования

$$x = \frac{a+b}{2} + \frac{b-a}{2}t.$$

Приведем для  $n = 1, 2, \dots, 5$  значения узлов, коэффициентов и остаточных членов в формуле

$$\int_{-1}^1 f(x) dx = \sum_{k=1}^n c_k f(x_k) + f^{(2n)}(\eta) \frac{2^{2n+1} (n!)^4}{[(2n)!]^3 (2n+1)}.$$

$n = 1$  :

$$x_1 = 0, \quad \frac{1}{2}c_1 = 1, \quad \Psi = \frac{1}{3}f''(\eta);$$

$n = 2$  :

$$\begin{aligned} -x_1 = x_2 = 0, 577\,350\,269\,189\,6258, \\ \frac{1}{2}c_1 = \frac{1}{2}c_2 = \frac{1}{2}, \quad \Psi = \frac{1}{135}f^{(4)}(\eta); \end{aligned}$$

$n = 3$  :

$$\begin{aligned} -x_1 = x_3 = 0, 774\,596\,669\,241\,4834, \quad x_2 = 0, \\ \frac{1}{2}c_1 = \frac{1}{2}c_3 = \frac{5}{18}, \quad \frac{1}{2}c_2 = \frac{4}{9}, \quad \Psi = \frac{1}{15\,750}f^{(6)}(\eta); \end{aligned}$$

$n = 4$  :

$$\begin{aligned} -x_1 = x_4 = 0, 861\,136\,311\,594\,0492, \\ -x_2 = x_3 = 0, 339\,981\,043\,584\,8646, \end{aligned}$$

$$\frac{1}{2}c_1 = \frac{1}{2}c_4 = 0,173\,927\,422\,568\,7284,$$

$$\frac{1}{2}c_2 = \frac{1}{2}c_3 = 0,326\,072\,577\,431\,2716,$$

$$\Psi = \frac{1}{3\,472\,875}f^{(8)}(\eta);$$

$n = 5 :$

$$-x_1 = x_5 = 0,906\,179\,845\,938\,6640,$$

$$-x_2 = x_4 = 0,538\,469\,310\,105\,6830, \quad x_3 = 0,$$

$$\frac{1}{2}c_1 = \frac{1}{2}c_5 = 0,118\,463\,442\,528\,0945,$$

$$\frac{1}{2}c_2 = \frac{1}{2}c_4 = 0,239\,314\,335\,249\,6832,$$

$$\frac{1}{2}c_3 = \frac{64}{225} = 0,284\,444\,444\,444\,4444,$$

$$\Psi = \frac{1}{1\,237\,732\,650}f^{(10)}(\eta).$$

## Упражнения

1. Вычислить следующие интегралы:

$$\int_0^1 \frac{dx}{1+x}, \quad \int_1^3 \frac{dx}{1+x}, \quad \int_0^1 \frac{dx}{1+x^2}, \quad \int_0^{\pi/2} \frac{\sin x}{x} dx,$$

$$\int_0^1 \frac{\ln(1+x)}{1+x^2} dx, \quad \int_0^1 \frac{\ln(1+x)}{x} dx,$$

с погрешностью, не превышающей  $10^{-4}$ , по приближенным формулам трапеций, Симпсона, Гаусса с одним и двумя узлами. Исследовать сходимость указанных методов на последовательности сеток  $(h, h/2, h/4)$ .

2. Провести аналогичное исследование для интеграла

$$\int_0^1 \sqrt{x} dx.$$

Результат обосновать.

3. Показать, что формула

$$\int_0^{+\infty} e^{-x} f(x) dx \approx \frac{2+\sqrt{2}}{4} f(2-\sqrt{2}) + \frac{2-\sqrt{2}}{4} f(2+\sqrt{2})$$

дает точные значения, если  $f(x)$  — многочлен степени не выше 4.

## ПРИБЛИЖЕННЫЕ МЕТОДЫ РЕШЕНИЯ УРАВНЕНИЙ

### § 4.1. ПОСТАНОВКА ЗАДАЧИ

Рассмотрим задачу нахождения корней уравнения вида

$$f(x) = 0, \quad (4.1)$$

где  $f(x)$  — функция действительного переменного.

К сожалению, в большинстве случаев формул для точного решения уравнения вида (4.1) не существует. Так, если  $f(x)$  — многочлен степени выше 4, то, согласно теореме Абеля, корни этого многочлена, вообще говоря, не могут быть выражены через его коэффициенты с помощью радикалов. Ненулевые корни простейшего трансцендентного уравнения  $\cos x - x = 0$  также не выражаются с помощью элементарных функций. Поэтому на практике часто прибегают к *итерационным методам* решения уравнений, суть которых состоит в построении числовой последовательности  $\{x_n\}$ , сходящейся к искомому корню  $x_*$  уравнения (4.1). Применение этих методов можно разбить на два этапа:

1) изучение расположения действительных корней и их отделение, то есть нахождение отрезков, на которых содержится ровно один корень;

2) построение итерационного процесса, позволяющего уточнить значение отыскиваемого корня с заданной точностью  $\varepsilon$ .

Рассмотрим каждый из этих этапов подробнее.

#### 4.1.1. ОТДЕЛЕНИЕ КОРНЕЙ. МЕТОД БИСЕКЦИИ

Отметим, что общих приемов отделения корней не существует. Для алгебраических уравнений эта задача может быть решена с помощью теоремы Штурма.

**Теорема 4.1** (Штурм). Пусть дано алгебраическое уравнение  $f(x) = 0$  степени  $n$ , не имеющее кратных корней; найдем производную  $f'(x) = f_1(x)$  и обозначим остаток от деления  $f(x)$  на  $f'(x)$ , взятый с обратным знаком, через  $f_2(x)$ ; остаток от деления  $f_1(x)$  на  $f_2(x)$  с обратным знаком — через  $f_3(x)$  и т. д., до тех пор, пока остаток не будет равен некоторой постоянной. Получим последовательность многочленов

$$f(x), f_1(x), \dots, f_n(x).$$

Тогда число действительных корней уравнения  $f(x) = 0$ , расположенных на отрезке  $[a, b]$ , равно разности между числом перемен знака нашей последовательности многочленов при  $x = a$  и числом перемен знака последовательности при  $x = b$ . (Доказательство этой теоремы приведено, например, в книге [7].)

Если уравнение имеет кратные корни, то, используя описанный алгоритм, можно получить последовательность многочленов  $f(x), f_1(x), \dots, f_m(x), 0$ , где  $f_m(x)$  не будет постоянным. Поделив все многочлены последовательности на  $f_m(x)$ , мы получим новую последовательность. Применяя к ней теорему Штурма, находим число корней уравнения  $f(x) = 0$ , расположенных на отрезке  $[a, b]$ , но без учета их кратности.

Для трансцендентных уравнений сколько-нибудь общего алгоритма отделения корней не существует. Приемы исследования функций подробно излагаются в курсе математического анализа, при этом достаточное условие наличия на отрезке хотя бы одного корня уравнения (4.1) формулируется в следующей теореме.

**Теорема 4.2** (Больцано — Коши). Пусть для непрерывной функции  $f(x)$ , определенной на отрезке  $[a, b]$ , выполняется неравенство  $f(a)f(b) < 0$ .

Тогда на интервале  $(a, b)$  существует корень уравнения (4.1). Если функция  $f(x)$  монотонна на отрезке  $[a, b]$ , то этот корень единственный. (Доказательство этой теоремы см., например, в книге [10].)

С теоремой Больцано — Коши тесно связан один из простейших итерационных методов — метод бисекции, называемый также методом деления отрезка пополам. Пусть



на отрезке  $[a, b]$  расположен единственный корень  $x_*$  уравнения (4.1), а  $f(a)$  и  $f(b)$  имеют разные знаки. Положим  $x_0 = (a + b)/2$  и вычислим  $f(x_0)$ . Очевидно,  $x_*$  принадлежит тому из интервалов  $(a, x_0)$ ,  $(x_0, b)$ , на границах которого функция  $f(x)$  имеет разные знаки. В качестве  $x_1$  берем середину этого интервала и повторяем описанный процесс до тех пор, пока на очередном шаге длина полученного интервала не станет меньше заданной точности  $\varepsilon$ . Тогда за приближенное значение корня принимается середина этого интервала. Формально метод бисекции может быть записан следующим образом:

$$x_{n+1} = x_n + \operatorname{sign} f(a) \cdot \operatorname{sign} f(x_n) \frac{b - a}{2^{n+2}}, \quad n = 0, 1, \dots,$$

где  $x_0 = (a + b)/2$ .

## § 4.2. ОСНОВЫ ТЕОРИИ ИТЕРАЦИОННЫХ МЕТОДОВ

Нахождение приближенного решения уравнения (4.1) с помощью итерационных методов начинается с замены этого уравнения эквивалентным уравнением

$$x = \Phi(x), \quad (4.2)$$

где

$$\Phi(x) = x + \tau(x)f(x), \quad (4.3)$$

причем функция  $\tau(x)$  не меняет знака на отрезке  $[a, b]$ , содержащем отыскиваемый корень (предполагается, что корни уравнения уже отделены). Далее функцию  $\Phi(x)$  будем считать непрерывной и достаточно гладкой.

Пусть  $x_*$  — корень уравнения (4.1) или, что то же самое, уравнения (4.2). Тогда

$$x_* = \Phi(x_*),$$

иными словами,  $x_*$  является *неподвижной точкой* функции  $\Phi$ . Для приближенного вычисления  $x_*$  образуем *итерационную последовательность*  $\{x_n\}$  по правилу

$$x_{n+1} = \Phi(x_n), \quad n = 0, 1, \dots, \quad (4.4)$$

где  $x_0$  — точка отрезка  $[a, b]$ , называемая *начальным приближением*. Итерационный метод вида (4.4) называют *методом простой итерации*. Заметим, что в форме метода простой итерации можно записать любой *одношаговый итерационный метод* (т. е. метод, в котором текущая итерация полностью определяется предыдущей).

Если последовательность  $\{x_n\}$  имеет предел  $\tilde{x}$ , причем  $\tilde{x} \in [a, b]$ , то, переходя в (4.4) к пределу при  $n \rightarrow \infty$ , ввиду непрерывности функции  $\Phi(x)$  получим  $\tilde{x} = \Phi(\tilde{x})$ , то есть последовательность  $\{x_n\}$  сходится к корню уравнения (4.2). Но по предположению уравнение (4.2) имеет на отрезке  $[a, b]$  единственный корень, поэтому  $\tilde{x} = x_*$ .

Разумеется, вид конкретного итерационного метода определяется выбором функции  $\tau(x)$  в формуле (4.3).

Естественно возникают два вопроса:

- 1) при каких условиях на  $\Phi(x)$  существует предел последовательности  $\{x_n\}$ ?
- 2) как быстро можно вычислить значение корня с заданной точностью  $\varepsilon$ ?

Рассмотрим каждый из этих вопросов отдельно.

#### 4.2.1. ПРИНЦИП СЖИМАЮЩИХ ОТОБРАЖЕНИЙ

Сформулируем условия на функцию  $\Phi$ , при которых итерационная последовательность  $\{x_n\}$  сходится. Напомним, что отображение  $\Phi$  называется *липшиц-непрерывным с постоянной  $q$*  на отрезке  $[a, b]$ , если для любых  $x, y \in [a, b]$  выполняется неравенство

$$|\Phi(x) - \Phi(y)| \leq q|x - y|.$$

При этом, если  $0 < q < 1$ , то отображение называется *сжимающим*. Нетрудно убедиться, что липшиц-непрерывное отображение является равномерно непрерывным.

Если функция  $\Phi(x)$  непрерывно дифференцируема на отрезке  $[a, b]$ , то в качестве постоянной  $q$  следует взять  $\max_{x \in [a, b]} |\Phi'(x)|$  (это видно из определения производной). Выполнение неравенства  $\max_{x \in [a, b]} |\Phi'(x)| < 1$  означает, что  $\Phi$  — сжимающее отображение.

Приведем одно весьма важное свойство сжимающих отображений.

**Теорема 4.3.** Пусть  $\Phi$  — отображение, являющееся сжимающим с постоянной  $q$  на отрезке  $U_r(x_0) = [x_0 - r, x_0 + r]$ , причем  $r \geq C/(1 - q)$ , где  $C = |\Phi(x_0) - x_0|$ .

Тогда отрезок  $U_r(x_0)$  переводится отображением  $\Phi$  в себя, т. е.  $\Phi(x) \in U_r(x_0)$  при всех  $x \in U_r(x_0)$ .

**Доказательство.** Нам нужно проверить, что при выполнении условий теоремы для любого  $x \in U_r(x_0)$  выполняется неравенство  $|\Phi(x) - x_0| \leq r$ . Но, так как отображение  $\Phi$  является сжимающим, то

$$\begin{aligned} |\Phi(x) - x_0| &\leq |\Phi(x) - \Phi(x_0)| + |\Phi(x_0) - x_0| \leq \\ &\leq q|x - x_0| + C \leq qr + (1 - q)r = r, \end{aligned}$$

что и требовалось доказать.

**Следствие 1.** Пусть  $\Phi$  — отображение, являющееся сжимающим с постоянной  $q$  на всей числовой оси. Тогда для любой точки  $x_0$  можно указать число  $r > 0$  такое, что отрезок  $U_r(x_0)$  переводится отображением  $\Phi$  в себя. В качестве  $r$  мы можем взять любое число, удовлетворяющее условию  $r \geq C/(1 - q)$ , где  $C = |\Phi(x_0) - x_0|$ .

**Следствие 2.** Пусть  $x_*$  — неподвижная точка отображения  $\Phi$ , гладкого на некотором отрезке  $U_r(x_*)$ , причем  $|\Phi'(x_*)| < 1$ . Тогда существует  $\delta > 0$  такое, что отрезок  $U_\delta(x_*)$  переводится отображением  $\Phi$  в себя.

**Доказательство.** Очевидно, найдутся числа  $q \in (0, 1)$  и  $\delta \in (0, r]$  такие, что  $|\Phi'(x)| \leq q$  для всех  $x \in U_\delta(x_*)$ . При этом  $\delta > |\Phi(x_*) - x_*|/(1 - q)$ , ибо  $\Phi(x_*) - x_* = 0$ . Таким образом, выполнены условия теоремы (4.3).

Теперь сформулируем и докажем утверждение, лежащее в основе теории итерационных методов.

**Теорема 4.4** (принцип сжимающих отображений). Пусть отображение  $\Phi$  является сжимающим с постоянной  $q$  на отрезке  $[a, b]$  и переводит этот отрезок в себя.

Тогда уравнение (4.2) имеет на отрезке  $[a, b]$  единственное решение  $x_*$ , при этом итерационная последовательность (4.4) сходится к  $x_*$  при любом начальном приближении  $x_0 \in [a, b]$ . Для погрешности имеет место оценка

$$|x_n - x_*| \leq q^n |x_0 - x_*|, \quad n = 1, 2, \dots$$

**Доказательство.** Так как отображение  $\Phi$  — сжимающее с постоянной  $q$ , то

$$|x_{j+1} - x_j| = |\Phi(x_j) - \Phi(x_{j-1})| \leq q |x_j - x_{j-1}|,$$

откуда

$$|x_{j+1} - x_j| \leq q^j |x_1 - x_0|, \quad j = 1, 2, \dots$$

Из последнего неравенства следует, что последовательность  $\{x_n\}$  является фундаментальной. В самом деле, для любого натурального  $p$  имеем

$$\begin{aligned} |x_{n+p} - x_n| &= \left| \sum_{j=1}^p (x_{n+j} - x_{n+j-1}) \right| \leq |x_1 - x_0| \sum_{j=1}^p q^{n+j-1} = \\ &= q^n \frac{1 - q^p}{1 - q} |x_1 - x_0| \leq \frac{q^n}{1 - q} |x_1 - x_0|. \end{aligned} \quad (4.5)$$

Правая часть неравенства (4.5) не зависит от  $p$ , поэтому из рассмотрения (4.5) при  $n \rightarrow \infty$  следует фундаментальность последовательности  $\{x_n\}$ , т. е. существование

$$\lim_{n \rightarrow \infty} x_n = x_* \in [a, b].$$

Так как отображение  $\Phi$  является сжимающим, то оно липшиц-непрерывно и, следовательно, непрерывно. Значит, как уже отмечалось ранее,  $x_*$  — корень уравнения (4.2).

Пусть  $x_{**} \in [a, b]$  — какой-либо корень уравнения (4.2). Тогда

$$|x_{**} - x_*| = |\Phi(x_{**}) - \Phi(x_*)| \leq q |x_{**} - x_*|.$$

Ввиду условия  $q < 1$  имеем  $x_{**} = x_*$ , то есть на отрезке  $[a, b]$  содержится единственный корень.

Оценим погрешность на  $n$ -й итерации. Для произвольного  $n$  имеет место неравенство

$$|x_n - x_*| = |\Phi(x_{n-1}) - \Phi(x_*)| \leq q|x_{n-1} - x_*|.$$

Применив последнее неравенство  $n$  раз, получим

$$|x_n - x_*| \leq q^n |x_0 - x_*|.$$

Чтобы проверить выполнение условий принципа сжимающих отображений, необходимо, прежде всего, определить области, в которых  $|\Phi'(x)| < 1$ . Отрезки, переводимые отображением  $\Phi(x)$  в себя, можно найти с помощью теоремы 4.3 и ее следствий. В частности, если корень уравнения (4.2) принадлежит области, в которой отображение  $\Phi$  сжимающее, то, согласно следствию 2 теоремы 4.3, существует окрестность этого корня, переводимая отображением  $\Phi$  в себя.

#### 4.2.2. СКОРОСТЬ СХОДИМОСТИ МЕТОДА

Пусть итерационная последовательность  $\{x_n\}$  имеет предел  $x_*$ . Говорят, что итерационный метод имеет  $k$ -й *порядок сходимости*, если существуют положительные постоянные  $c_1, c_2$  такие, что

$$|x_{n+1} - x_*| \leq c_2 |x_n - x_*|^k \quad (4.6)$$

при условии  $|x_n - x_*| \leq c_1$ .

На практике часто встречаются методы, имеющие первый порядок сходимости (иначе говоря, *линейно сходящиеся*). Для них оценка (4.6) имеет вид

$$|x_n - x_*| \leq q|x_{n-1} - x_*|, \quad n = 0, 1, \dots,$$

где  $q < 1$ . Отсюда следует уже знакомая нам формула:

$$|x_n - x_*| \leq q^n |x_0 - x_*|. \quad (4.7)$$

Таким образом, итерационные методы, удовлетворяющие условиям принципа сжимающих отображений имеют по крайней мере первый порядок сходимости.

Очевидно, формула (4.7) неприменима в практических вычислениях, ибо содержит в правой части неизвестную величину  $x_*$ . Легко получить более грубую оценку, содержащую в правой части только известные величины:

$$|x_n - x_*| \leq q^n |b - a|, \quad (4.8)$$

где  $[a, b]$  — отрезок, содержащий предел  $x_*$  и начальное приближение  $x_0$ . В соответствии с последней формулой линейно сходящиеся методы называют еще *методами, сходящимися со скоростью геометрической прогрессии*, а число  $q$  — ее *знаменателем*.

Из оценки (4.8) следует, что для гарантированного достижения точности  $\varepsilon$  требуется сделать  $\left\lceil \frac{\ln(\varepsilon/|b-a|)}{\ln q} \right\rceil + 1$  итераций (в данном случае квадратные скобки обозначают целую часть числа).

Если итерационная последовательность  $\{x_n\}$  образована по формуле (4.4), в которой отображение  $\Phi$  — сжимающее с постоянной  $q$ , то можно получить еще одну оценку сходимости. Так как

$$\begin{aligned} |x_0 - x_*| &= |x_0 - \Phi(x_*)| \leq |x_0 - \Phi(x_0)| + |\Phi(x_0) - \Phi(x_*)| \leq \\ &\leq |x_0 - \Phi(x_0)| + q|x_0 - x_*|, \end{aligned}$$

то

$$|x_0 - x_*| \leq \frac{1}{1-q} |\Phi(x_0) - x_0|.$$

Подставляя последнее неравенство в формулу (4.7), имеем

$$|x_n - x_*| \leq \frac{q^n}{1-q} |\Phi(x_0) - x_0|.$$

С помощью этой оценки также можно заранее подсчитать число итераций, требуемое для достижения точности  $\varepsilon$ .

**Задача 4.1.** Пусть в формуле (4.4) отображение  $\Phi$  — сжимающее с постоянной  $q$ . Доказать, что имеет место оценка

$$|x_n - x_*| \leq \frac{q}{1-q} |x_n - x_{n-1}|.$$

Теперь сформулируем и докажем теорему, позволяющую определять скорость сходимости методов, имеющих порядок сходимости выше первого.

**Теорема 4.5.** Пусть  $x_*$  — неподвижная точка отображения  $\Phi \in C^m$ , причем

$$\Phi'(x_*) = \Phi''(x_*) = \dots = \Phi^{(m-1)}(x_*) = 0, \quad \Phi^{(m)}(x_*) \neq 0.$$

Тогда итерационный метод (4.4) в некоторой окрестности  $U_r(x_*)$  имеет  $m$ -й порядок сходимости. Кроме того, если

$$|x_0 - x_*| < 1, \quad \frac{M_m}{m!} |x_0 - x_*| = \omega < 1, \quad (4.9)$$

где  $M_m = \max_{x \in U_r(x_*)} |\Phi^{(m)}(x)|$ , то имеет место оценка погрешности:

$$|x_n - x_*| \leq \omega^{\frac{m^n - 1}{m - 1}}. \quad (4.10)$$

**Доказательство.** По формуле Тейлора

$$\begin{aligned} \Phi(x) - \Phi(x_*) &= (x - x_*)\Phi'(x_*) + \frac{(x - x_*)^2}{2!}\Phi''(x_*) + \dots \\ &\dots + \frac{(x - x_*)^{(m-1)}}{(m-1)!}\Phi^{(m-1)}(x_*) + \frac{(x - x_*)^m}{m!}\Phi^{(m)}(\xi). \end{aligned}$$

С учетом условий теоремы последнее равенство запишется в виде

$$\Phi(x) - x_* = \frac{(x - x_*)^m}{m!}\Phi^{(m)}(\xi),$$

откуда после подстановки  $x = x_{n-1}$  получим

$$x_n - x_* = \frac{(x_{n-1} - x_*)^m}{m!}\Phi^{(m)}(\xi_n).$$

Значит,

$$|x_n - x_*| \leq \frac{M_m}{m!} |x_{n-1} - x_*|^m. \quad (4.11)$$

Последовательно используя неравенство (4.11)  $n$  раз, мы получим следующую оценку:

$$\begin{aligned} |x_n - x_*| &\leq \left( \frac{M_m}{m!} \right)^{1+m+m^2+\dots+m^{n-1}} |x_0 - x_*|^{m^n} = \\ &= \left( \frac{M_m}{m!} \right)^{\frac{m^n - 1}{m - 1}} |x_0 - x_*|^{m^n} = \end{aligned}$$

$$= \left( \frac{M_m}{m!} |x_0 - x_*| \right)^{\frac{m^n - 1}{m - 1}} |x_0 - x_*|^{\frac{m^{n+1} - 2m^n + 1}{m - 1}}.$$

Отсюда при выполнении условий (4.9) следует неравенство (4.10).

Оценка (4.10) устанавливает возможность очень быстрой сходимости итерационного процесса, но для ее применения требуется выполнение неравенств (4.9). Это означает, что начальное приближение  $x_0$  должно быть достаточно близко к искомому корню. Последнее обстоятельство является существенным недостатком методов высокого порядка сходимости, поэтому иногда несколько начальных членов последовательности  $\{x_n\}$  вычисляется с помощью всегда сходящегося метода (например метода бисекции), после чего, когда погрешность станет достаточно малой, применяется какой-либо метод высокого порядка.

Отметим, что для ряда методов (например для метода секущих, описанного в п. 4.3.3) вычисления в окрестности кратного корня сопровождаются потерей значащих цифр. Это приводит к так называемой «разболтке» счета или даже к резкому возрастанию погрешности вплоть до переполнения разрядной сетки ЭВМ (см. рис. 4.1). Избежать этого позволяет прием Гарвика, суть которого состоит в следующем.

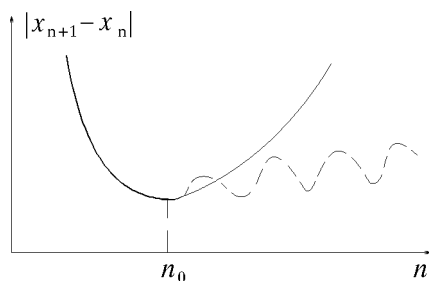


Рис. 4.1. Прием Гарвика

Выбрав не очень малое  $\varepsilon$ , проводят итерации до выполнения условия  $|x_{n+1} - x_n| < \varepsilon$ , а затем продолжают расчет до тех пор, пока значения  $|x_{n+1} - x_n|$  убывают. Первое же возрастание обычно означает начало «разболтки»; тогда расчет



прекращают и последнюю итерацию не используют. Уточнить результат можно с помощью более надежного метода первого порядка, например метода бисекции.

В заключение дадим геометрическую интерпретацию итерационного метода (4.4). Рис. 4.2, *а* соответствует случаю, когда  $0 < \Phi'(x) \leq q < 1$ , рис. 4.2, *б* — когда  $-1 < -q \leq \Phi'(x) < 0$ .

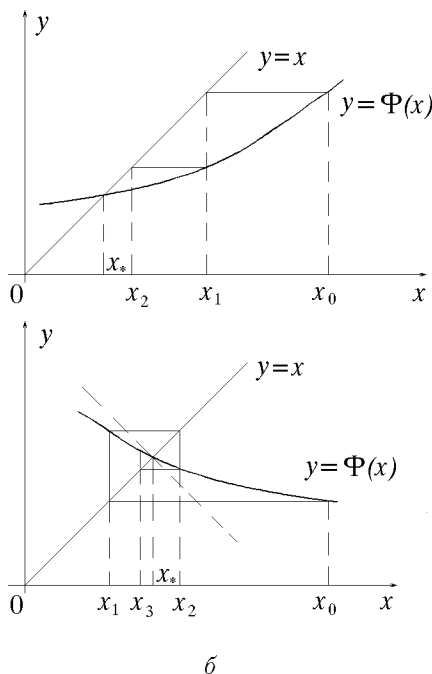


Рис. 4.2. Монотонная и немонотонная сходимость итераций

Из рисунка видно, что в случае  $0 < \Phi'(x) \leq q < 1$  имеет место монотонная сходимость последовательности  $\{x_n\}$  к корню, причем все члены последовательности лежат с той стороны  $x_*$ , с которой расположено начальное приближение  $x_0$ . Если же  $-1 < -q \leq \Phi'(x) < 0$ , то члены последовательности располагаются поочередно с разных сторон от точки  $x = x_*$ , то есть выполняется неравенство  $|x_n - x_*| \leq |x_n - x_{n-1}|$ , позволяющее легко оценивать точность, достигнутую на очередном шаге.

**Задача 4.2.** Аналитически доказать утверждения, сформулированные в последнем абзаце.

## § 4.3. ПРИМЕРЫ ИТЕРАЦИОННЫХ МЕТОДОВ

### 4.3.1. МЕТОД РЕЛАКСАЦИИ

Напомним, что по построению  $\Phi(x) = x + \tau(x)f(x)$ . Пусть  $\tau(x) = \tau = \text{const} \neq 0$ . Тогда итерационный метод, называемый *методом релаксации*, запишется в виде

$$x_{n+1} = x_n + \tau f(x_n), \quad n = 0, 1, 2, \dots,$$

при этом  $\tau$  иногда называют *релаксационным параметром*.

Проверим выполнение условий принципа сжимающих отображений. Будем считать, что уравнение  $f(x) = 0$  имеет корень  $x_*$ , локализованный на отрезке  $[a, b]$ . Очевидно,  $\Phi'(x) = 1 + \tau f'(x)$ , т. е. отображение  $\Phi$  является сжимающим, если для всех  $x \in [a, b]$  выполняется условие

$$-2 < \tau f'(x) < 0. \quad (4.12)$$

Не уменьшая общности, далее в этом пункте будем считать, что  $f'(x) < 0$  и, кроме того, в некоторой окрестности корня  $0 < m_1 < |f'(x)| < M_1$ . Тогда неравенство (4.12) имеет место при

$$\tau \in (0, 2/M_1). \quad (4.13)$$

Так как  $x_*$  — неподвижная точка отображения  $\Phi$ , то, согласно следствию 2 теоремы 4.3, при выполнении условия (4.13) существует некоторая окрестность  $U_r(x_*)$ , переводимая отображением  $\Phi$  в себя, т. е. в этой окрестности метод релаксации сходится.

Теперь займемся нахождением оптимального значения параметра  $\tau$ . Естественным критерием оптимальности будет служить скорость сходимости метода, возрастающая с уменьшением постоянной  $q = \max |\Phi'(x)|$ . Для метода простой итерации  $|\Phi'(x)| = |1 + \tau f'(x)|$ , поэтому из условия  $0 < m_1 < |f'(x)| < M_1$ , записанного в виде  $-M_1 < f'(x) < -m_1 < 0$ , следует, что

$$q(\tau) = \max\{|1 - \tau m_1|, |1 - \tau M_1|\}.$$

Нам требуется найти  $\tau$ , при котором значение  $q$  минимально.

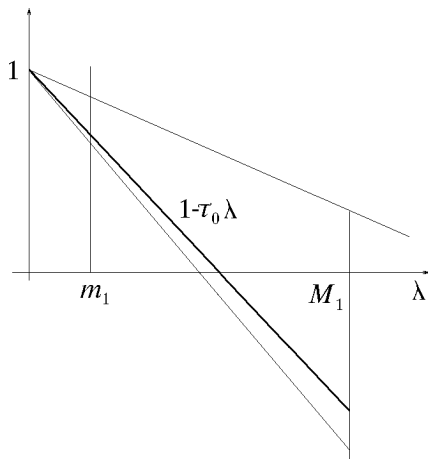


Рис. 4.3. Оптимизация релаксационного параметра

Заметим, что график функции  $1 - \tau\lambda$  аргумента  $\lambda$  представляет собой прямую (рис. 4.3). Если  $0 < \tau \leq 1/M_1$ , то функция  $1 - \tau\lambda$  неотрицательна на отрезке  $[m_1, M_1]$  и  $q(\tau) = 1 - \tau m_1$ . Если же  $\tau > 1/M_1$ , то величина  $1 - \tau M_1$  становится отрицательной, ее модуль возрастает с ростом  $\tau$ , причем для некоторого  $\tau = \tau_0$  выполняется равенство

$$1 - \tau_0 m_1 = -(1 - \tau_0 M_1). \quad (4.14)$$

Если  $0 < \tau < \tau_0$ , то  $q(\tau) = 1 - \tau m_1 > 1 - \tau_0 m_1 = q(\tau_0)$ ; если  $\tau_0 < \tau$ , то  $q(\tau) = \tau M_1 - 1 > \tau_0 M_1 - 1 = q(\tau_0)$ . Значит,  $q(\tau)$  минимально при  $\tau = \tau_0$ . Из (4.14) следует, что

$$\tau_0 = \frac{2}{M_1 + m_1},$$

откуда

$$q(\tau_0) = \frac{M_1 - m_1}{M_1 + m_1}.$$

### 4.3.2. МЕТОД НЬЮТОНА

Пусть уравнение  $f(x) = 0$  имеет на отрезке  $[a, b]$  единственный корень  $x_*$  и, кроме того, на этом отрезке функция  $f$  дважды непрерывно дифференцируема, а ее производная  $f'(x)$  не обращается в нуль. Тогда, положив в формуле (4.3)  $\tau(x) = -\frac{1}{f'(x)}$ , мы получим уравнение вида (4.2):

$$x = x - \frac{f(x)}{f'(x)} \equiv \Phi(x).$$

Так как

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

то  $\Phi'(x_*) = 0$ . Нетрудно убедиться, что, вообще говоря,  $\Phi''(x_*) \neq 0$ , т. е. согласно теореме 4.5 итерационный метод

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad n = 0, 1, \dots, \quad (4.15)$$

называемый *методом Ньютона*, имеет второй порядок сходимости.

Заметим, что значение  $x_{n+1}$  есть абсцисса точки пересечения с осью  $Ox$  касательной, проведенной в точке  $(x_n, f(x_n))$  к графику функции  $f(x)$  (см. рис. 4.4). Именно поэтому метод Ньютона называют также *методом касательных*.

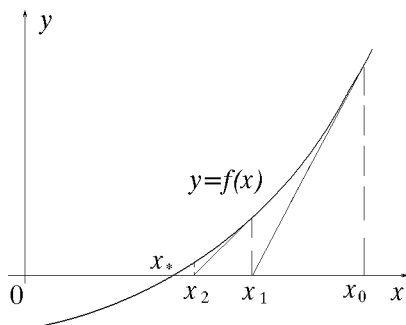


Рис. 4.4. Метод Ньютона

Для оценки скорости сходимости метода Ньютона разложим функцию  $f(x)$  в ряд Тейлора в окрестности точки  $x_n$ :

$$0 = f(x_*) = f(x_n) + f'(x_n)(x_* - x_n) + \frac{1}{2}f''(\xi)(x_* - x_n)^2,$$

где  $\xi$  лежит между  $x_n$  и  $x_*$ . Значит,

$$\frac{f(x_n)}{f'(x_n)} = x_n - x_* - \frac{1}{2} \frac{f''(\xi)}{f'(x_n)} (x_* - x_n)^2.$$

Подставляя это выражение в формулу (4.15), получим

$$x_{n+1} - x_* = x_n - x_* - \frac{f(x_n)}{f'(x_n)} = \frac{1}{2} \frac{f''(\xi)}{f'(x_n)} (x_* - x_n)^2.$$

Отсюда

$$|x_{n+1} - x_*| \leq \frac{M_2}{2m_1} |x_n - x_*|^2,$$

где  $m_1 = \min_{x \in [a, b]} |f'(x)|$ , а  $M_2 = \max_{x \in [a, b]} |f''(x)|$ .

Конечно, полученная оценка может гарантировать квадратичную сходимость метода лишь начиная с того номера  $n$ , для которого выполняется условие  $\frac{M_2}{2m_1} |x_n - x_*| < 1$ . Начальное приближение  $x_0$  желательно выбирать так, чтобы имело место неравенство

$$f(x_0)f''(x_0) > 0,$$

ибо в противном случае, как видно из рис. 4.5, слишком большая начальная погрешность  $|x_0 - x_*|$  может привести к тому, что итерационный процесс не будет сходящимся.

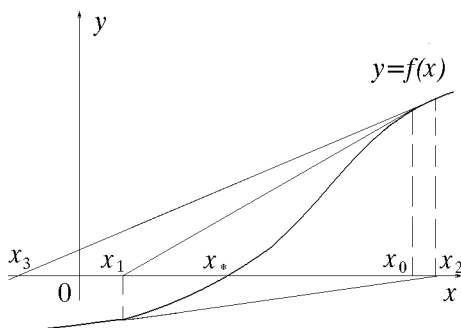


Рис. 4.5. Расходимость метода Ньютона

Достаточный признак сходимости метода Ньютона сформулирован в следующей теореме (предполагается, что  $x_*$  — единственный корень уравнения (4.1) на отрезке  $[a, b]$ ).

**Теорема 4.6.** Пусть для всех  $x \in [a, b]$  выполнено неравенство

$$f'(x)f''(x) > 0.$$

Тогда итерационная последовательность, построенная по методу Ньютона при  $x_0 = b$ , монотонно убывает и сходится к  $x_*$ .

Если же для всех  $x \in [a, b]$  выполнено неравенство

$$f'(x)f''(x) < 0,$$

то итерационная последовательность, построенная по методу Ньютона при  $x_0 = a$ , монотонно возрастает и сходится к  $x_*$ .

**Задача 4.3.** Доказать теорему 4.6.

**Указание.** Воспользоваться формулой конечных приращений Лагранжа.

Один из случаев монотонной сходимости метода Ньютона был продемонстрирован на рис. 4.4.

Опишем простой прием контроля точности в методе Ньютона. Так как, согласно формуле Тейлора,  $f(x_n) = f(x_n) - f(x_*) = (x_n - x_*)f'(\xi)$ , то

$$|x_n - x_*| \leq \frac{|f(x_n)|}{m_1}. \quad (4.16)$$

Эта оценка позволяет на каждом шаге следить за достигнутой точностью.

Наконец заметим, что вблизи кратного корня может наблюдаться «разболтка» численного решения ввиду потери точности из-за ошибок округления при вычислении отношения двух малых величин. Поэтому следует применять прием Гарвика. Это относится и к другим методам высокого порядка.

#### 4.3.2. МОДИФИКАЦИИ МЕТОДА НЬЮТОНА

Опишем ряд итерационных методов, получающихся путем упрощения метода Ньютона. В этом пункте мы будем полагать, что функция  $f$  и ее производные обладают всеми свойствами, которые необходимы для применения метода Ньютона (см. начало пункта 4.3.2).

Чтобы избежать в формуле (4.15) многократного вычисления производной  $f'(x)$ , используют *модифицированный метод Ньютона*:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)}, \quad n = 0, 1, 2, \dots$$

Очевидно, этот метод является частным случаем ( $\tau = -1/f'(x_0)$ ) метода релаксации, условия сходимости которого приведены в п. 4.3.1. Для их выполнения необходимо, в частности, чтобы начальное приближение  $x_0$  удовлетворяло требованию

$$|f'(x_0)| > \frac{M_1}{2}.$$

Последнее неравенство сразу следует из условия (4.13) (напомним, что  $M_1 = \max |f'(x)|$ ).

Таким образом, модифицированный метод Ньютона предъявляет меньше требований к выбору начального приближения  $x_0$ , но имеет лишь первый порядок сходимости.

Как видно из рис. 4.6, на первом шаге модифицированного метода Ньютона строится касательная к графику функции  $f(x)$  в точке  $x_0$ , а на последующих шагах — секущие, параллельные этой касательной.

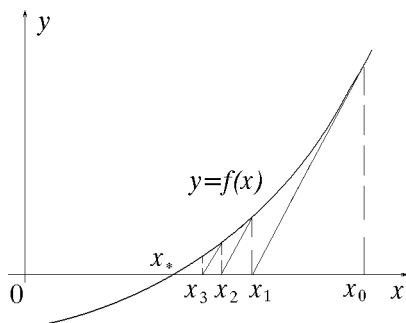


Рис. 4.6. Модифицированный метод Ньютона

Если в методе Ньютона (4.15) аппроксимировать  $f'(x_n)$  конечно-разностным отношением  $\frac{f(x_n) - f(x_0)}{x_n - x_0}$ , то получим *метод хорд*:

$$x_{n+1} = x_n - \frac{x_n - x_0}{f(x_n) - f(x_0)} f(x_n), \quad n = 1, 2, \dots$$

В некоторых книгах этот метод называют «методом секущих», но, как мы увидим чуть ниже, последнее название логичнее использовать для другого метода.

В качестве  $x_0$  выбирают точку, в которой  $f(x_0)f''(x_0) > 0$ . Начальным приближением фактически является некоторая точка  $x_1$ , удовлетворяющая условию  $f(x_1)f(x_0) < 0$ . Последующие итерации осуществляются обычным способом.

Чтобы определить условия сходимости метода хорд, заметим, что в данном случае

$$\Phi(x) = x - \frac{x - x_0}{f(x) - f(x_0)} f(x) = \frac{x_0 f(x) - x f(x_0)}{f(x) - f(x_0)},$$

то есть

$$\begin{aligned} \Phi'(x_*) &= \\ &= \frac{(x_0 f'(x_*) - f(x_0))(f(x_*) - f(x_0)) - f'(x_*)(x_0 f(x_*) - x_* f(x_0))}{(f(x_*) - f(x_0))^2} = \\ &= \frac{f(x_0) + (x_* - x_0)f'(x_*)}{f(x_0)}. \end{aligned}$$

С другой стороны, по формуле Тейлора

$$f(x_0) + (x_* - x_0)f'(x_*) = \frac{(x_0 - x_*)^2}{2!} f''(\xi).$$

Тем самым

$$\Phi'(x_*) = \frac{(x_0 - x_*)^2}{2} \frac{f''(\xi)}{f(x_0)}.$$

Выбирая  $x_0$  достаточно близким к  $x_*$ , мы можем добиться выполнения неравенства  $\Phi'(x) \leq q < 1$  в некоторой окрестности точки  $x_*$ , что обеспечит сходимость метода хорд с первым порядком, при этом для оценки точности может быть использовано неравенство (4.16).



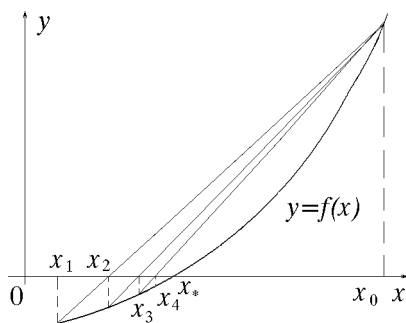


Рис. 4.7. Метод хорд

Геометрическая трактовка метода хорд заключается в том, что значение  $x_{n+1}$  есть абсцисса точки пересечения прямой, проходящей через точки  $(x_0, f(x_0))$  и  $(x_n, f(x_n))$ , с осью  $Ox$ , то есть из точки  $(x_0, f(x_0))$  проводится семейство хорд графика функции  $f(x)$  (рис. 4.7). Иными словами, на каждом шаге за приближенное значение  $x_{n+1}$  принимается корень интерполяционного многочлена первой степени, построенного по значениям  $f(x)$  в точках  $x_0$  и  $x_n$ . Именно поэтому данный метод иногда называют *методом линейной интерполяции*.

При аппроксимации в методе Ньютона производной  $f'(x_n)$  конечно-разностным отношением  $\frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$  получается *метод секущих*:

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

Этот метод отличается от ранее изученных тем, что на каждом шаге для нахождения  $x_{n+1}$  используется не только значение  $x_n$ , но и значение  $x_{n-1}$ . Методы, обладающие таким свойством, называются *двухшаговыми*. Они требуют задания двух начальных приближений:  $x_0$  и  $x_1$ .

Отметим еще одну любопытную особенность метода секущих: его порядок сходимости (см. формулу (4.6)) есть иррациональное число  $k \approx 1,6180$ , являющееся положительным корнем уравнения  $k^2 - k - 1 = 0$  и называемое *золотым сечением*. Это число имеет важное значение в теории чисел Фибоначчи.

Контроль точности в рассматриваемом методе может быть осуществлен с помощью формулы (4.16).

Геометрически метод секущих состоит в том, что через точки  $(x_{n-1}, f(x_{n-1}))$  и  $(x_n, f(x_n))$  проводится прямая, абсцисса точки пересечения которой с осью  $Ox$  является новым приближением  $x_{n+1}$  (рис. 4.8). Пределом построенного семейства секущих является касательная к графику функции  $f(x)$  в точке  $x_*$ . Заметим, что в основе метода секущих, как и метода хорд, лежит идея линейной интерполяции.

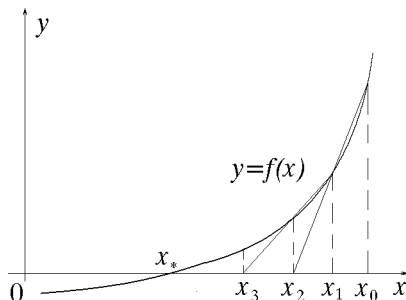


Рис. 4.8. Метод секущих

#### 4.3.4. МЕТОДЫ ЧЕБЫШЕВА ВЫСОКИХ ПОРЯДКОВ СХОДИМОСТИ

Приведем разработанный П.Л. Чебышевым алгоритм построения итерационных методов, имеющих сколь угодно высокий порядок сходимости. В основе этого алгоритма лежит разложение в ряд Тейлора функции, обратной к функции  $f(x)$ .

Пусть уравнение (4.1) имеет на отрезке  $[a, b]$  единственный корень  $x_*$ , причем функция  $y = f(x)$  на этом отрезке достаточно гладкая и строго монотонная (последнее означает, что  $f'(x) \neq 0$ , т. е.  $[a, b]$  однозначно отображается в некоторый отрезок  $[c, d]$ ). Тогда существует функция  $x = g(y)$ , обратная к  $y = f(x)$ , определенная на  $[c, d]$  и имеющая такой же порядок гладкости, как  $f(x)$ . Очевидно, имеет место равенство  $x \equiv g[f(x)]$ , последовательное дифференцирование

которого дает нам систему линейных уравнений для нахождения производных функции  $g(y)$ , причем матрица этой системы имеет треугольный вид:

$$\begin{aligned} g'[f(x)]f'(x) &= 1, \\ g'[f(x)]f''(x) + g''[f(x)](f'(x))^2 &= 0, \\ g'[f(x)]f'''(x) + 3g''[f(x)]f'(x)f''(x) + g'''[f(x)](f'(x))^3 &= 0 \\ &\dots\dots\dots \end{aligned} \quad (4.17)$$

Отсюда следует, что

$$g'[f(x)] = \frac{1}{f'(x)}, \quad g''[f(x)] = -\frac{f''(x)}{(f'(x))^3}$$

и т. д.

Приступая непосредственно к описанию алгоритма Чебышева, заметим, что  $x_* = g(0)$ . Рассмотрим произвольную точку  $y \in [c, d]$ . Согласно формуле Тейлора,

$$g(0) = g(y) + \sum_{k=1}^r (-1)^k \frac{g^{(k)}(y)}{k!} y^k + (-1)^{r+1} \frac{g^{(r+1)}(\eta)}{(r+1)!} y^{r+1}$$

( $\eta$  лежит между 0 и  $y$ ). Переходя к переменной  $x$ , получим

$$x_* = x + \sum_{k=1}^r (-1)^k \frac{g^{(k)}[f(x)]}{k!} (f(x))^k + (-1)^{r+1} \frac{g^{(r+1)}(\eta)}{(r+1)!} (f(x))^{r+1}, \quad (4.18)$$

причем производные функции  $g[f(x)]$  могут быть найдены из системы (4.17).

Введем обозначение

$$\Phi_r(x) = x + \sum_{k=1}^r (-1)^k \frac{g^{(k)}[f(x)]}{k!} (f(x))^k.$$

Тогда, очевидно, уравнение

$$x = \Phi_r(x)$$

имеет корень  $x_*$ . Кроме того,

$$\Phi_r^{(l)}(x_*) = 0, \quad l = 1, 2, \dots, r. \quad (4.19)$$

Таким образом, итерационный метод

$$x_{n+1} = \Phi_r(x_n), \quad n = 0, 1, \dots,$$

вследствие теоремы 4.5 имеет  $(r+1)$ -й порядок сходимости.

**Задача 4.4.** Доказать равенство (4.19).

Заметим, что решение задачи 4.4 требует весьма громоздких выкладок, поэтому порядок сходимости метода проще найти непосредственно, т. е. без использования теоремы 4.5. Для этого подставим в (4.18)  $x = x_{n-1}$ . Тогда

$$x_* - x_n = (-1)^{r+1} \frac{g^{(r+1)}[f(\xi)]}{(r+1)!} (f(x_{n-1}))^{r+1},$$

где  $\xi$  лежит между  $x_*$  и  $x_{n-1}$ . Пусть  $L = \max_{x \in [a, b]} |f'(x)|$ ,

$M_{r+1} = \max_{x \in [a, b]} |g^{(r+1)}[f(x)]|$ . Так как

$$|f(x_{n-1})| = |f(x_{n-1}) - f(x_*)| = |f'(\eta)| |x_{n-1} - x_*| \leq L |x_{n-1} - x_*|,$$

то

$$\begin{aligned} |x_n - x_*| &\leq \frac{M_{r+1} L^{r+1}}{(r+1)!} |x_{n-1} - x_*|^{r+1} \leq \\ &\leq \left( \frac{M_{r+1} L^{r+1}}{(r+1)!} \right)^{1+(r+1)+(r+1)^2+\dots+(r+1)^{n-1}} |x_0 - x_*|^{(r+1)^n} = \\ &= \left( \frac{M_{r+1} L^{r+1}}{(r+1)!} |x_0 - x_*| \right)^{\frac{(r+1)^n - 1}{r}} |x_0 - x_*|^{\frac{(r+1)^n (r-1) + 1}{r}}. \end{aligned}$$

Итак, если  $|x_0 - x_*| < 1$  и  $\frac{M_{r+1} L^{r+1}}{(r+1)!} |x_0 - x_*| = \omega < 1$ , то имеет место оценка

$$|x_n - x_*| \leq \omega^{\frac{(r+1)^n - 1}{r}},$$

указывающая на очень быструю сходимость метода. Обратите внимание, что эта оценка совершенно аналогична оценке (4.10), полученной при доказательстве теоремы 4.5.

Теперь приведем конкретный вид формул, получающихся при  $r = 1$  и  $r = 2$ . Нетрудно убедиться, что

$$\Phi_1(x) = x - \frac{f(x)}{f'(x)},$$

т. е. с помощью алгоритма Чебышева может быть получен метод Ньютона. Далее,

$$\Phi_2(x) = x - \frac{f(x)}{f'(x)} - \frac{f''(x)f^2(x)}{2(f'(x))^3}$$

и

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} - \frac{f''(x_n)f^2(x_n)}{2(f'(x_n))^3}. \quad (4.20)$$

Заметим, что иногда *методом Чебышева* называют именно метод (4.20), имеющий третий порядок сходимости, ибо методы, соответствующие  $r \geq 3$ , на практике применяют весьма редко ввиду сложности вычислений и необходимости знать производные высокого порядка.

#### 4.3.5. МЕТОД ПАРАБОЛ

Как отмечено в п. 4.3.3, метод хорд является, по сути, методом линейной интерполяции. Общая же идея так называемых *интерполяционных методов* состоит в том, что нахождение корней уравнения (4.1) заменяется нахождением корней интерполяционного многочлена, построенного для  $f(x)$ . Если интерполяционный многочлен имеет вторую степень, то полученный метод называется *методом парабол*. Выпишем его формулы. Пусть известны приближения  $x_{n-2}$ ,  $x_{n-1}$ ,  $x_n$ . Построим интерполяционный многочлен Ньютона

$$\begin{aligned} L_2(x) = & f(x_n) + (x - x_n)f(x_n, x_{n-1}) + \\ & + (x - x_n)(x - x_{n-1})f(x_n, x_{n-1}, x_{n-2}). \end{aligned}$$

Приравнявая его нулю, получим квадратное уравнение

$$az^2 + bz + c = 0,$$

где

$$z = x - x_n, \quad a = f(x_n, x_{n-1}, x_{n-2}),$$

$$b = f(x_n, x_{n-1}) + (x_n - x_{n-1})f(x_n, x_{n-1}, x_{n-2}), \quad c = f(x_n).$$

Тот из двух корней квадратного уравнения, который меньше по модулю, определяет новое приближение  $x_{n+1} = x_n + z$ .

Метод парабол является *трехшаговым*, и для начала расчета требуется задать три начальных приближения  $x_0, x_1, x_2$ . Метод сходится с порядком  $k \approx 1,84$ .

Отметим, что корни решаемого квадратного уравнения могут быть и комплексными. Тем самым метод парабол позволяет получать и комплексные корни уравнения (4.1), хотя начальное приближение может быть и вещественным. Это свойство делает метод парабол весьма удобным для поиска всех корней многочлена. Как правило, сначала ищут корни, наименьшие по модулю, ибо при нахождении больших по модулю корней возрастает вычислительная погрешность. Очередной найденный корень  $x_*$  удаляют путем алгебраического деления на  $(x - x_*)$ , при этом в качестве следующих начальных приближений берут  $x_0 = 0, x_1 = -1, x_2 = 1$ .

Интерполяционные методы с использованием многочленов порядка выше 2 не получили распространения, так как существенно возрастает число арифметических действий на каждой итерации, а порядок сходимости не достигает 2.

#### 4.3.6. МЕТОД ЭЙТКЕНА УСКОРЕНИЯ СХОДИМОСТИ

Опишем предложенный А. Эйткеном прием, позволяющий получить из двух итерационных методов одного и того же порядка (эти методы, вообще говоря, могут совпадать) новый метод более высокого порядка.

**Теорема 4.7.** Пусть итерационные методы

$$x_{n+1}^{(1)} = \Phi_1(x_n^{(1)}), \quad x_{n+1}^{(2)} = \Phi_2(x_n^{(2)}), \quad n = 0, 1, \dots \quad (4.21)$$

имеют порядок  $r \in \mathbf{N}$  и сходятся к  $x = x_*$ . Построим функцию  $\Phi(x)$  следующего вида:

$$\Phi(x) \equiv \frac{x\Phi_1[\Phi_2(x)] - \Phi_1(x)\Phi_2(x)}{x - \Phi_1(x) - \Phi_2(x) + \Phi_1[\Phi_2(x)]}.$$

Тогда итерационный метод

$$x_{n+1} = \Phi(x_n), \quad n = 0, 1, \dots \quad (4.22)$$

имеет порядок выше  $r$ , если выполнено условие

$$(\Phi'_1(x_*) - 1)(\Phi'_2(x_*) - 1) \neq 0. \quad (4.23)$$

**Доказательство.** Выберем произвольное число  $\tilde{x}$  и образуем две последовательности:  $y_n^{(i)} = x_n^{(i)} - \tilde{x} \quad (i = 1, 2)$ , каждая из которых сходится к  $x_* - \tilde{x}$ . Так как

$$y_n^{(i)} + \tilde{x} = x_n^{(i)} = \Phi_i(x_{n-1}^{(i)}) = \Phi_i(y_{n-1}^{(i)} + \tilde{x}),$$

то нетрудно убедиться, что указанные последовательности получаются при решении уравнений

$$y = \Phi_i(y + \tilde{x}) - \tilde{x}.$$

При  $\tilde{x} = x_*$  последовательности примут вид  $\eta_n^{(i)} = x_n^{(i)} - x_*$ , они получаются из уравнений

$$\eta = \Psi_i(\eta) \equiv \Phi_i(\eta + x_*) - x_*,$$

а их члены представляют собой отклонения  $x_n^{(i)}$  от точного значения корня  $x_*$ .

Последовательности (4.21) имеют порядок  $r$ , поэтому мы можем записать разложение

$$\Psi_i(\eta) = \psi_r^{(i)} \eta^r + \psi_{r+1}^{(i)} \eta^{r+1} + \dots, \quad i = 1, 2. \quad (4.24)$$

Далее, из определений функций  $\Phi(x)$  и  $\Psi_i(\eta)$  следует, что

$$\begin{aligned} \Phi(\eta + x_*) &= \frac{(\eta + x_*)\Phi_1[\Phi_2(\eta + x_*)] - \Phi_1(\eta + x_*)\Phi_2(\eta + x_*)}{\eta + x_* - \Phi_1(\eta + x_*) - \Phi_2(\eta + x_*) + \Phi_1[\Phi_2(\eta + x_*)]} = \\ &= \frac{(\eta + x_*)(\Psi_1[\Psi_2(\eta)] + x_*) - (\Psi_1(\eta) + x_*)(\Psi_2(\eta) + x_*)}{\eta - \Psi_1(\eta) - \Psi_2(\eta) - x_* + \Psi_1[\Psi_2(\eta)] + x_*} = \\ &= \frac{\eta\Psi_1[\Psi_2(\eta)] - \Psi_1(\eta)\Psi_2(\eta) + x_*(\eta - \Psi_1(\eta) - \Psi_2(\eta) + \Psi_1[\Psi_2(\eta)])}{\eta - \Psi_1(\eta) - \Psi_2(\eta) + \Psi_1[\Psi_2(\eta)]}, \end{aligned}$$

то есть

$$\Phi(\eta + x_*) - x_* = \frac{\eta\Psi_1[\Psi_2(\eta)] - \Psi_1(\eta)\Psi_2(\eta)}{\eta - \Psi_1(\eta) - \Psi_2(\eta) + \Psi_1[\Psi_2(\eta)]}.$$

Подставим в последнюю формулу разложения (4.24):

$$\Phi(\eta + x_*) - x_* =$$

$$\begin{aligned}
&= \frac{\eta(\psi_r^{(1)}(\psi_r^{(2)}\eta^r + \dots)^r + \dots) - (\psi_r^{(1)}\eta^r + \dots)(\psi_r^{(2)}\eta^r + \dots)}{\eta - (\psi_r^{(1)}\eta^r + \dots) - (\psi_r^{(2)}\eta^r + \dots) + (\psi_r^{(1)}(\psi_r^{(2)}\eta^r + \dots)^r + \dots)} = \\
&= \frac{(\psi_r^{(1)}\psi_r^{(2)r}\eta^{r^2+1} + \dots) - (\psi_r^{(1)}\psi_r^{(2)}\eta^{2r} + \dots)}{\eta - (\psi_r^{(1)} + \psi_r^{(2)})\eta^r + \psi_r^{(1)}\psi_r^{(2)r}\eta^{r^2} + \dots}.
\end{aligned}$$

Если  $r \geq 2$ , то наименьшая степень  $\eta$  в числителе равна  $2r$ , а в знаменателе 1. Тем самым разложение в ряд Тейлора функции  $\Phi(\eta + x_*)$  по степеням  $\eta$  начинается с  $\eta^{2r-1}$ , т. е.  $\Phi^{(l)}(x_*) = 0$  при  $l = 1, 2, \dots, 2r - 2$ . Значит, согласно теореме 4.5, итерационный метод (4.22) имеет порядок не ниже  $2r - 1$ .

Если же  $r = 1$ , то наименьшая степень  $\eta$  в числителе не меньше 3, ибо члены степени 2 взаимно уничтожаются. В знаменатель  $\eta$  входит с коэффициентом

$$\begin{aligned}
1 - \psi_1^{(1)} + \psi_1^{(2)} + \psi_1^{(1)}\psi_1^{(2)} &= (\psi_1^{(1)} - 1)(\psi_1^{(2)} - 1) = \\
&= (\Phi_1'(x_*) - 1)(\Phi_2'(x_*) - 1).
\end{aligned}$$

Таким образом, при выполнении условия (4.23) в знаменателе присутствует член первой степени, т. е. разложение функции  $\Phi(\eta + x_*)$  по степеням  $\eta$  начинается по крайней мере с  $\eta^2$ . Значит,  $\Phi'(x_*) = 0$  и итерационный метод (4.22) имеет порядок не ниже 2.

Рассмотрим важный случай, когда  $\Phi_1(x) = \Phi_2(x) = \tilde{\Phi}(x)$ . Тогда

$$\Phi(x) \equiv \frac{x\tilde{\Phi}[\tilde{\Phi}(x)] - \tilde{\Phi}^2(x)}{x - 2\tilde{\Phi}(x) + \tilde{\Phi}[\tilde{\Phi}(x)]}. \quad (4.25)$$

Однако для построения итерационного процесса в данном случае нет нужды находить явный вид функции  $\Phi(x)$ . Действительно, задав  $x_0$ , находим

$$x_1 = \tilde{\Phi}(x_0), \quad x_2 = \tilde{\Phi}(x_1),$$

после чего вычисляем  $x_3$  с помощью формулы (4.25):

$$x_3 = \frac{x_0x_2 - x_1^2}{x_0 - 2x_1 + x_2} = x_0 - \frac{(x_1 - x_0)^2}{(x_2 - x_1) - (x_1 - x_0)}. \quad (4.26)$$

Положив

$$\Delta x_i = x_{i+1} - x_i, \quad \Delta^2 x_i = \Delta x_{i+1} - \Delta x_i,$$



представим (4.26) в виде

$$x_3 = x_0 - \frac{(\Delta x_0)^2}{\Delta^2 x_0}.$$

Затем описанный трехшаговый цикл повторяется:

$$\begin{aligned}x_{3i+1} &= \tilde{\Phi}(x_{3i}), \\x_{3i+2} &= \tilde{\Phi}(x_{3i+1}), \\x_{3(i+1)} &= x_{3i} - \frac{(\Delta x_{3i})^2}{\Delta^2 x_{3i}}.\end{aligned}$$

В заключение отметим, что по функции  $\tilde{\Phi}(x)$  описанным способом можно построить итерационный процесс еще более высокого порядка и т. д.

#### 4.3.7. СРАВНЕНИЕ СВОЙСТВ МЕТОДОВ

При решении на ЭВМ уравнения вида (4.1) неизбежно встает вопрос: какой метод выбрать? Разумеется, важной характеристикой метода является скорость сходимости, однако в методах высокого порядка для выполнения одной итерации требуется провести больше арифметических операций, что неизбежно сказывается на скорости решения задачи. Кроме того, эти методы обычно требуют достаточно точного выбора начального приближения, вследствие чего начальные итерации приходится выполнять с помощью метода первого порядка (например бисекции).

Далее, в методах Ньютона, Чебышева и др. требуется вычислять производные высокого порядка, что сопряжено с известными сложностями. Наконец, «разболтка» вблизи кратного корня ввиду потери точности из-за ошибок округления тоже характерна для методов, имеющих высокий порядок сходимости, поэтому в таких случаях заключительные итерации иногда выполняются с помощью методов первого порядка, устойчивых к «разболтке». Например, в методе парабол при последовательном вычислении корней многочлена целесообразно удалять очередной найденный корень  $x_*$  делением на  $x - x_*$ . Вследствие этого после удаления нескольких корней происходит накопление ошибки, которую можно существенно уменьшить, применяя надежный метод бисекции.

Обобщая вышесказанное, подчеркнем, что правильный выбор методов и их разумное комбинирование является важным элементом численного решения уравнений. Умение сделать такой выбор, приобретаемое по мере накопления опыта, свидетельствует о профессионализме математика-вычислителя.

#### § 4.4. РЕШЕНИЕ СИСТЕМ НЕЛИНЕЙНЫХ УРАВНЕНИЙ

#### 4.4.1. ПОСТАНОВКА ЗАДАЧИ

Рассмотрим систему нелинейных уравнений

[illegible]

где  $f_i$ ,  $i = 1, 2, \dots, m$ , — функции действительных переменных  $x_1, x_2, \dots, x_m$ . Систему (4.27) удобно рассматривать как операторное уравнение в пространстве  $\mathbf{R}^m$ . Это уравнение имеет вид

$$F(x) = 0, \quad (4.28)$$

где  $x=(x_1, x_2, \dots, x_m)^T \in H$ ,  $F(x)=(f_1(x), f_2(x), \dots, f_m(x))^T$  — отображение из  $\mathbf{R}^m$  в  $\mathbf{R}^m$ .

Для решения системы (4.28) обычно используются итерационные методы, которые заключаются в построении последовательности векторов  $\{x^n\}$  (обратите внимание: в отличие от скалярного случая номер итерации пишется *вверху*), сходящейся к вектору  $x_*$  — точному решению уравнения (4.28). Напомним два определения сходимости последовательности векторов. *Покоординатная сходимость* означает, что  $x_i^n \rightarrow x_{*i}$  при  $n \rightarrow \infty$  для всех  $i = 1, 2, \dots, m$ . *Сходимость по норме*  $\|\cdot\|_*$  означает, что  $\|x^n - x_*\|_* \rightarrow 0$  при  $n \rightarrow \infty$ .

Рассмотрим следующие нормы векторов:

$$\|x\|_1 = \sum_{i=1}^m |x_i|, \quad \|x\|_2 = \sqrt{\sum_{i=1}^m x_i^2}, \quad \|x\|_\infty = \max_{1 \leq i \leq m} |x_i|.$$

Из курса алгебры известно, что в конечномерном пространстве покоординатная сходимость эквивалентна сходимости в любой норме. Однако при решении задачи о нахождении корня уравнения (4.28) с заранее заданной точностью  $\varepsilon$  необходимо уточнить, в какой именно норме должна достигаться указанная точность (обычно подразумевается сходимость в норме  $\|\cdot\|_\infty$ , то есть покоординатная сходимость).

В дальнейшем нам также понадобится понятие *матричной нормы*, подчиненной заданной векторной норме. По определению,

$$\|A\|_* = \max_{\|x\|_* = 1} \|Ax\|_*.$$

**Задача 4.5.** Доказать, что

$$\|A\|_1 = \max_{1 \leq j \leq m} \sum_{i=1}^m |a_{ij}|, \quad \|A\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^m |a_{ij}|, \quad (4.29)$$

где  $a_{ij}$  — элементы матрицы  $A$ .

Если в одношаговый итерационный метод для решения системы (4.28) новая итерация  $x^{n+1}$  входит линейно, то он может быть записан в так называемой *канонической форме*:

$$B_{n+1} \frac{x^{n+1} - x^n}{\tau_{n+1}} + F(x^n) = 0, \quad n = 1, 2, \dots, \quad (4.30)$$

где  $x^0$  — заданное начальное приближение,  $\tau_{n+1}$  — числовые параметры,  $B_{n+1}$  — невырожденная матрица порядка  $m \times m$ . Заметим, что выражение (4.30) аналогично канонической форме одношагового итерационного метода для решения систем линейных уравнений [6].

Из (4.30) следует, что  $x^{n+1}$  строится по известному  $x^n$  путем решения системы линейных алгебраических уравнений

$$B_{n+1}x^{n+1} = g(x^n), \quad (4.31)$$

где  $g(x^n) = B_{n+1}x^n - \tau_{n+1}F(x^n)$ , при этом для решения системы (4.31) могут использоваться как прямые, так и итерационные методы. В последнем случае итерации, предназначенные для решения системы (4.31), называют *внутренними итерациями*, а итерации (4.30) — *внешними итерациями*.

Если  $B_{n+1} = E$  для всех  $n \in \mathbf{N}$ , то метод (4.30) называется *явным*, в противном случае —  *неявным*. Если ни  $B$ , ни  $\tau$  не зависят от номера итерации  $n$ , то метод называется *стационарным*.

Определение *порядка сходимости* совершенно аналогично соответствующему определению для скалярного случая, данному в п. 4.2.2 (разумеется, следует заменить  $|\cdot|$  на  $\|\cdot\|$ ).

Отметим, что нахождение приближенных решений системы (4.27) является весьма сложной задачей, особенно при достаточно больших  $m$ . Сложности порождаются, в частности, отсутствием сколько-нибудь общих приемов отделения корней, невозможностью дать наглядную геометрическую интерпретацию итерационного процесса, трудоемкостью вычислений и т. п. Решение системы нелинейных уравнений в каждом конкретном случае может потребовать специального громоздкого исследования. Ввиду этого говорить об универсальности алгоритмов решения многомерных систем не приходится. При решении практических задач приходится заниматься адаптацией того или иного численного метода к рассматриваемому классу задач.

#### 4.4.2. СХОДИМОСТЬ СТАЦИОНАРНЫХ МЕТОДОВ

Стационарный метод вида (4.30) можно записать как

$$x^{n+1} = \Phi(x^n), \quad (4.32)$$

при этом исходное уравнение (4.28) примет вид

$$x = \Phi(x),$$

где  $\Phi(x) = x - \tau B^{-1}F(x)$ .

Таким образом, решение уравнения (4.28) сводится к решению уравнения  $x_* = \Phi(x_*)$ , т. е. к отысканию неподвижных точек оператора  $\Phi$ . С такой задачей мы уже сталкивались в скалярном случае. Она решалась на основе принципа сжимающих отображений, сформулированного и доказанного в теореме 4.4. Аналогичное утверждение верно и для многомерного случая.

Говорят, что оператор  $\Phi$  является *сжимающим с постоянной  $q$  на множестве  $K \subseteq \mathbf{R}^m$* , если существует число  $q \in (0, 1)$  такое, что для любых  $x, y \in K$  выполняется неравенство

$$\|\Phi(x) - \Phi(y)\| \leq q\|x - y\|.$$

Далее будем считать, что  $K$  — выпуклая область.

Очевидно, данное определение зависит от выбора нормы, ибо оператор, являющийся в некоторой норме  $\|\cdot\|_*$  сжимающим, может не быть таковым в другой норме.

**Задача 4.6.** Построить оператор, сжимающий в норме  $\|\cdot\|_\infty$ , но не являющийся таковым в норме  $\|\cdot\|_1$ .

**Задача 4.7.** Построить оператор, сжимающий в норме  $\|\cdot\|_1$ , но не являющийся таковым в норме  $\|\cdot\|_\infty$ .

Выясним условия, достаточные для того, чтобы оператор  $\Phi$  был сжимающим в некоторой норме. Предположим, что  $\Phi \in C^1$ ,  $x, y \in K$ . Применяя формулу Тейлора для функции многих переменных, имеем

$$\Phi(y) = \Phi(x) + \Phi'(\xi)(y - x), \quad (4.33)$$

где  $\xi = x + \theta(y - x)$ ,  $0 < \theta < 1$ ,  $\Phi'(\xi)$  — матрица Якоби отображения  $\Phi$ , имеющая следующий вид:

$$\Phi'(\xi) = \begin{pmatrix} \frac{\partial \Phi_1(\xi)}{\partial x_1} & \cdots & \frac{\partial \Phi_1(\xi)}{\partial x_m} \\ \cdots & \cdots & \cdots \\ \frac{\partial \Phi_m(\xi)}{\partial x_1} & \cdots & \frac{\partial \Phi_m(\xi)}{\partial x_m} \end{pmatrix}.$$

Из (4.33) следует, что

$$\|\Phi(y) - \Phi(x)\| \leq \|\Phi'(\xi)\| \|y - x\|,$$

Таким образом, если существует  $q \in (0, 1)$  такое, что для любого  $\xi \in K$  выполняется оценка

$$\|\Phi'(\xi)\| \leq q, \quad (4.34)$$

то оператор  $\Phi$  является сжимающим с постоянной  $q$  на множестве  $K$ .

Для проверки оценки (4.34) на практике удобно использовать формулы (4.29). При этом следует отметить, что нахождение точных границ множества  $K$  является, как правило,

весьма трудной задачей. Не менее сложно определить, переводит ли отображение  $\Phi$  некоторую многомерную область в себя. Поэтому, хотя доказательство многомерных аналогов теорем 4.3 и 4.4 не составляет особого труда, мы сформулируем принцип сжимающих отображений для систем уравнений без использования понятия «области, переводимой в себя».

**Теорема 4.8** (принцип сжимающих отображений для систем уравнений). Пусть оператор  $\Phi$ , определенный на множестве

$$U_r(a) = \{x \in \mathbf{R}^m : \|x - a\| \leq r\},$$

является сжимающим на этом множестве с коэффициентом сжатия  $q$ , причем

$$\|\Phi(a) - a\| \leq (1 - q)r.$$

Тогда в  $U_r(a)$  оператор  $\Phi$  имеет единственную неподвижную точку  $x_*$  и итерационный метод (4.32) сходится к  $x_*$  при любом  $x^0 \in U_r(a)$ , причем для погрешности справедливы оценки

$$\|x^n - x_*\| \leq q^n \|x^0 - x_*\|,$$

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

**Следствие.** Если  $\|\Phi'(x_*)\| < 1$ , то метод (4.32) сходится по крайней мере с первым порядком.

**Задача 4.8.** Доказать теорему 4.8.

Еще раз напомним, что определение сжимающего оператора зависит от выбора нормы. Но если с помощью принципа сжимающих отображений удастся показать сходимость итерационного метода (4.31) в некоторой норме  $\|\cdot\|_*$ , то, вследствие эквивалентности норм в конечномерных пространствах, метод будет сходиться в любой норме. Поэтому достаточно установить, что условия теоремы 4.8 выполнены хотя бы для одной нормы.

#### 4.4.3. ПРИМЕРЫ ИТЕРАЦИОННЫХ МЕТОДОВ

**Пример 1.** *Метод релаксации* представляет собой явный стационарный одношаговый метод. Он получается, если в формуле (4.30) положить  $B_{n+1} = E$ ,  $\tau_{n+1} = \tau$ , т. е.

$$x^{n+1} = x^n - \tau F(x^n).$$

Согласно следствию из теоремы 4.8, метод релаксации сходится с первым порядком, если  $\|E - \tau F'(x_*)\| < 1$ , где  $F'(x)$  — матрица Якоби отображения  $F(x)$ .

**Пример 2.** *Метод Ньютона* для системы нелинейных уравнений можно получить с помощью алгоритма Чебышева (см. п. 4.3.4), примененного к функции  $m$  переменных. Указанный алгоритм приводит к системе

$$F'(x^n)(x^{n+1} - x^n) + F(x^n) = 0.$$

Таким образом, чтобы получить метод Ньютона, в (4.30) полагают  $B_{n+1} = F'(x^n)$ ,  $\tau_{n+1} = 1$ , при этом необходима обратимость матриц  $F'(x^n)$ . Другая форма записи метода:

$$x^{n+1} = x^n - (F'(x^n))^{-1}F(x^n).$$

Если начальное приближение  $x^0$  достаточно близко к корню, то метод Ньютона сходится со вторым порядком (доказательство см., например, в [3]).

**Пример 3.** *Модифицированный метод Ньютона* применяют, чтобы избежать обращения матрицы  $F'(x)$  на каждой итерации:

$$F'(x^0)(x^{n+1} - x^n) + F(x^n) = 0.$$

Тем самым построен стационарный итерационный метод, имеющий первый порядок сходимости.

В заключение рассмотрим два итерационных метода, в которые новая итерация  $x^{n+1}$  входит нелинейно. Имеет место идейное сходство этих методов с одноименными методами решения систем линейных уравнений.

**Пример 4.** *Нелинейный метод Якоби* для системы (4.27) имеет вид

$$f_i(x_1^n, x_2^n, \dots, x_{i-1}^n, x_i^{n+1}, x_{i+1}^n, \dots, x_m^n) = 0, \quad i = 1, 2, \dots, m.$$

Таким образом, для нахождения  $x^{n+1}$  необходимо решить  $m$  скалярных уравнений. Их независимость делает метод Якоби удобным для реализации на многопроцессорных машинах. Разумеется, каждое скалярное уравнение можно решать своим итерационным методом, наиболее подходящим для этого уравнения.

**Пример 5.** *Нелинейный метод Зейделя* имеет вид

$$f_i(x_1^{n+1}, x_2^{n+1}, \dots, x_i^{n+1}, x_{i+1}^n, \dots, x_m^n) = 0$$

и сводится к последовательному решению выписанных уравнений относительно неизвестной  $x_i^{n+1}$ ,  $i = 1, 2, \dots, m$ . В отличие от метода Якоби, в методе Зейделя найденные в ходе текущей итерации значения компонент вектора  $x^{n+1}$  подставляются во все последующие уравнения.

## Упражнения

1. Используя метод Ньютона для решения уравнения  $x^2 - N = 0$ , построить алгоритм приближенного вычисления  $\sqrt{N}$ .

2. Исследовать поведение итерационной последовательности метода Ньютона для уравнения  $x^3 - x = 0$  в зависимости от выбора начального приближения  $x_0$ . Определить, для каких  $x_0$  при отсутствии вычислительной погрешности последовательность не сходится.

3. Решить приближенно с точностью до  $10^{-7}$  систему уравнений  $2x^3 - y^2 - 1 = 0$ ;  $xy^3 - y - 4 = 0$  методом Ньютона, взяв начальное приближение из области, ограниченной прямыми  $y = 0$ ,  $y = x$ ,  $x = 0,5$ . Пронаблюдать поведение итераций, определив порядок сходимости.

4. Рассмотрим два итерационных метода. В одном из них внешние итерации осуществляются по Ньютону, а внутреннее — по Зейделю. В другом для построения  $x^{n+1}$  используется метод Зейделя, а каждое из скалярных уравнений решается методом Ньютона. Показать, что описанные методы не совпадают.



## СПИСОК ЛИТЕРАТУРЫ

1. *Ахмеров Р.Р., Коробицына Ж.Л., Слепцов А.Г.* Основы численного анализа в задачах. Новосибирск: Изд-во НГУ, 1994.
2. *Бараннин В.Б., Шапеев В.П.* Введение в численный анализ. Новосибирск: Изд-во НГУ, 1997.
3. *Батталов Н.С.* Численные методы. М.: Наука, 1973.
4. *Березин И.С., Жидков Н.П.* Методы вычислений. Т. I. М.: Физматгиз, 1959. Т. II. М.: Физматгиз, 1962.
5. *Калиткин Н.Н.* Численные методы. М.: Наука, 1978.
6. *Коновалов А.Н.* Введение в вычислительные методы линейной алгебры. Новосибирск: Наука, 1993.
7. *Курош А.Г.* Курс высшей алгебры. М.: Физматгиз, 1959.
8. *Марчук Г.И.* Методы вычислительной математики. М.: Наука, 1989.
9. *Никольский С.М.* Квадратурные формулы. М.: Наука, 1979.
10. *Никольский С.М.* Курс математического анализа. Т. I. М.: Наука, 1990.
11. Пособие по практике на ЭВМ. Новосибирск: НГУ, 1980.
12. *Самарский А.А., Гулин А.В.* Численные методы. М.: Наука, 1989.
13. *Турчак Л.И.* Основы численных методов. М.: Наука, 1987.
14. *Яненко Н.Н., Шокин Ю.И.* Численный анализ. Теория приближения функций. Новосибирск: НГУ, 1980.

# ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

- Авост** 10  
**Аппроксимация** 15  
**Аппроксимирующая функция** 15
- Базисный многочлен Лагранжа** 17
- Взвешенное среднеквадратичное** 46
- Значащая цифра** 13  
— — верная 13  
— — сомнительная 13
- Интерполирование** 15  
— кусочно-полиномиальное 36  
**Интерполяционный многочлен** 16  
— — Лагранжа 17  
— — Ньютона 20  
— — Эрмита 23
- Итерационная последовательность** 73  
**Итерационный метод** 71  
— — бисекции 72  
— — двухшаговый 89  
— — Зейделя нелинейный 104  
— — интерполяционный 93  
— — касательных 84  
— — линейно сходящийся 77  
— — неявный 100  
— — Ньютона 84, 103  
— — модифицированный 87, 103  
— — одношаговый 74  
— — парабол 93  
— — простой итерации 74  
— — секущих 89  
— — релаксации 82, 103  
— — стационарный 100  
— — трехшаговый 94  
— — хорд 88  
— — Чебышева 93  
— — явный 100  
— — Якоби нелинейный 103
- Каноническая форма одношагового итерационного метода** 99  
**Квадратурная формула** 49  
Квадратурная формула Гаусса 62  
— — интерполяционного типа 50  
— — Ньютона — Котеса 50  
— — прямоугольников 51  
— — — составная 52  
— — симметричная 57  
— — Симпсона (парабол) 54  
— — — составная 55  
— — трапеций 52  
— — — составная 53  
**Коэффициент квадратурной формулы** 49  
**Кратность узла интерполирования** 23
- Мантисса числа** 9  
**Машинная бесконечность** 10  
**Машинный ноль** 9  
**Метод наименьших квадратов** 32  
— неопределенных коэффициентов 27  
— Рунге 59  
— Эйткена 94
- Начальное приближение** 75  
**Неподвижная точка функции** 74
- Округление** 10  
**Остаточный член интерполяционной формулы** 21  
**Отображение липшиц-непрерывное** 74  
— сжимающее 74
- Приближение функций** 15  
**Прием Гарвика** 82  
**Погрешность абсолютная** 10  
— — предельная 10  
— вычислительная 8  
— квадратурной формулы 49  
— математической модели 7  
— метода 8  
— округления 8  
— относительная 10  
— — предельная 11  
— неустраняемая 8  
— решения полная 9

- Позиционная система счисления 9
- Порядок сходимости итерационного метода 74, 100
  - точности квадратурной формулы 50
  - формулы численного дифференцирования 26
  - числа 9
- Разность вперед 25
  - назад 25
  - разделенная 18
  - центральная 26
- Разряд числа 9
- Релаксационный параметр 82
- Сетка 25, 49
- Сжимающий оператор 100
- Сплайн 36
  - кубический 37
  - сглаживающий 44
- Среднеквадратичное уклонение 35
- Таблицы значений узлов и коэффициентов формул Гаусса 68
- Узел интерполирования 15
  - квадратурной формулы 49
- Численное дифференцирование 25
- Число с плавающей точкой 9
  - с фиксированной точкой 9
- Экстраполирование 22