НИЖЕГОРОДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМ.Н.И.ЛОБАЧЕВСКОГО

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

УЧЕБНО-НАУЧНЫЙ И ИННОВАЦИОННЫЙ КОМПЛЕКС
"НОВЫЕ МНОГОФУНКЦИОНАЛЬНЫЕ МАТЕРИАЛЫ И
НАНОТЕХНОЛОГИИ"

Фаддеев М.А., Марков К.А.

ЧИСЛЕННЫЕ МЕТОДЫ

(Учебное пособие)

Мероприятие 1.2. Совершенствование образовательных технологий, укрепление материально-технической базы учебного процесса

Учебные дисциплины: «Информатика», «Численные методы»

Специальности, направления: «Физика», «Нанотехнология», «Нанотехнология в электронике», «Микроэлектроника и полупроводниковые приборы», «Электроника и наноэлектроника»

ННГУ, 2010

В любой науке столько истины,
сколько в ней математики.

Иммануил Кант

ВВЕДЕНИЕ

Все естественные науки используют вычисления в своей практике. Еще в V веке до нашей эры философ Пифагор, которого можно назвать основоположником математической физики, утверждал, что всё сущее управляется числами. Все значительные этапы в развитии физики сопровождались разработкой новых разделов математики. Величайший физик и математик Исаак Ньютон при формулировке основ механики, которые были изложены в его знаменитых «Математических началах натуральной философии», опубликованных в 1687 году, разработал дифференциальное и интегральное исчисление. Примерно 50 лет спустя Леонард Эйлер создал вариационное исчисление, решая актуальные задачи строительной механики. Создание молекулярно-кинетической теории теплоты происходило вместе с развитием теории вероятностей. Классическая электродинамика дала мощный импульс исследованиям дифференциальных уравнений в частных производных. Исследование устойчивости динамических систем привело великого математика и физика Анри Пуанкаре к разработке теории бифуркаций. Квантовая механика базируется на теории операторов в гильбертовом пространстве. Аналогичное влияние решения физических проблем на развитие новых разделов математики наблюдается в областях квантовой хромодинамики, нелинейных колебаний, теории единого поля и т.д.

Во все времена практика решения прикладных задач демонстрировала недостаточность имеющихся в наличии аналитических методов математики и необходимость разработки численных методов. Например, невозможность при интегрировании представить первообразную функцию в аналитическом виде приводит к использованию квадратурных формул приближенного вычисления определенного интеграла. Приближенный характер результатов численных методов не является принципиальным препятствием к их использованию, так как в физике применяются те численные методы, погрешность которых может быть сделана ниже приемлемой точности результата данной задачи.

Принципиально новая фаза в развитии численных методов наступила с началом широкого использования компьютерной техники в физических исследованиях. Важнейшее достоинство компьютеров – высокая скорость выполнения математических операций – позволило в исторически короткий срок создать новые сравнительно простые и эффективные алгоритмы. Современные компьютеры за доли секунды выполняют огромное количество арифметических действий с многоразрядными числами, что обеспечивает требуемую точность результатов. При этом отпала необходимость в использовании множества хитроумных приемов вычислений «с помощью карандаша и бумаги», каждый из которых разрабатывался для довольно узкого класса задач.

Повсеместное распространение персональных компьютеров обеспечивает широкую доступность численных методов. При этом у начинающих исследователей часто возникают трудности выбора конкретного метода для решения поставленной задачи.

Количество численных методов, разработанных к настоящему времени, огромно. Одни обладают большей общностью, другие имеют весьма специальное назначение и используются для узкого круга физических задач. Таким образом, начинающие исследователи нуждаются в информации, позволяющей выбрать оптимальный численный метод.

Положение с учебной литературой по численным методам в настоящее время не является благополучным. Сохранились в небольшом количестве старые учебники: толстые, добротные, изобилующие вышеупомянутыми специальными приемами вычислений «с помощью бумаги и карандаша». Большая часть объема таких томов посвящена доказательствам, которые существенны для математиков, но для физиков-экспериментаторов имеют чисто академический интерес. К тому же изучение таких учебников требует солидной математической подготовки, и поэтому тяжело для студентов первого и второго курсов, которым уже приходится обращаться к некоторым численным методам.

С другой стороны, существуют многочисленные методические пособия из нескольких страниц, содержащие перечисление названий методов и список формул. Учебно-методическая ценность таких опусов близка к нулю. Наконец, имеются современные учебные пособия (как правило, переводы зарубежных изданий), в которых изложены численные методы, адаптированные к современным компьютерам. К сожалению, такие учебники издаются недостаточными тиражами,

имеют высокую розничную цену и поэтому недоступны подавляющему числу студентов.

По мнению авторов, которые преподают курс численных методов на физическом факультете ННГУ, студенты нуждаются в учебном пособии, в котором должны быть изложены самые элементарные методы, необходимые начинающему физику. Этот минимальный набор должен, прежде всего, содержать простейшие методы приближенного вычисления значений функций: интерполяцию и аппроксимацию. В набор должны быть включены методы численного интегрирования и дифференцирования. Многие важные практические задачи физики требуют применения численных методов решения нелинейных уравнений и систем линейных уравнений.

Для логической связности изложения оказалось необходимым включить в пособие разделы вычисления конечных разностей, определителей матриц и обратных матриц.

При отборе материала авторы ограничивались самыми элементарными методами, которые обеспечивают приемлемую погрешность при реализации этих методов на персональных компьютерах.

Настоящее издание не содержит методов решения дифференциальных уравнений, так как, по мнению авторов, эта тема требует подробной разработки и изложения в отдельном учебном пособии. Также авторы отказались от систематического изложения в данном издании теории погрешностей, так как по этой теме уже опубликовано большое количество удачных учебников и методических указаний.

Предлагаемое учебное пособие ориентировано в первую очередь на студентов, магистрантов и аспирантов физического факультета. Не возбраняется использование данного издания школьниками, занимающимися научной работой на физическом факультете в рамках научного общества учащихся.

Авторы надеются, что настоящее учебное пособие будет полезным и для студентов других факультетов физико-математических и естественно-научных направлений.

В процессе подготовки данного издания авторам оказывали помощь многие сотрудники физического факультета ННГУ. Авторы благодарны всем, кто взял на себя труд ознакомится с рукописью и дать полезные советы. Особую благодарность авторы выражают профессорам В.Р. Фидельману, Е.В. Чупрунову, доцентам В.А. Бурдову, Н.С. Будникову,

Г.М. Максимовой, О.А. Морозову, ассистенту М.О. Марычеву и инженеру Е.А. Конькову. Авторы признательны рецензентам данной работы – сотрудникам кафедры теоретической механики механико-математического факультета ННГУ, в первую очередь декану механико-математического факультета, д.ф.-м.н. А.К. Любимову, доцентам А.Ф. Ляхову и Д.Т. Чекмареву, а также доценту кафедры математического обеспечения ЭВМ факультета вычмслительной математики и кибернетики, к.ф.-м.н. В.А.Гришагину за многочисленные ценные рекомендации.

Глава 1. КОНЕЧНЫЕ РАЗНОСТИ

Различные численные методы используют величины, которые называются конечными разностями.

Пусть $y = f(x)$ – некоторая заданная функция. Обозначим $\Delta x \equiv h$ постоянное конечное приращение аргумента (шаг).

Определение. Величина

$$\Delta y \equiv \Delta f(x) = f(x+\Delta x) - f(x) \equiv f(x+h) - f(x) \tag{1.1}$$

называется ***первой конечной разностью*** функции $y = f(x)$.

При заданном постоянном шаге h конечная разность является функцией аргумента x.

Конечные разности высших порядков определяются индуктивно:

$$\Delta^n y = \Delta(\Delta^{n-1} y), \quad n = 2, 3, \ldots \tag{1.2}$$

В частности, согласно определению (1.2), вторая конечная разность представляется как разность первых конечных разностей

$$\Delta^2 y = \Delta(\Delta y) = \Delta\,(f(x+h) - f(x))$$

и может быть выражена в следующем виде через значения исходной функции:

$$\Delta^2 y = f(x+2\Delta x) - 2f(x+\Delta x) + f(x) \tag{1.3}$$

Пример 1. Рассмотрим степенную функцию $y = x^3$ и последовательно вычислим ее конечные разности различных порядков:

$$\Delta y = 3x^2 h + 3xh^2 + h^3, \quad \Delta^2 y = 6xh^2 + 6h^3, \quad \Delta^3 y = 6h^3,$$

$$\Delta^n y = 0 \quad (n \geq 3).$$

Приведенный пример демонстрирует, что конечная разность произвольного порядка является, вообще говоря, функцией аргумента x и содержит постоянный параметр h. Численное значение любой конечной разности относится к конкретной точке x (конечно, кроме тех случаев, когда конечная разность является константой).

Пример 1 со степенной функцией показывает, что конечные разности такой функции постепенно понижают свой показатель степени каждый раз на единицу. Так как показатель этой степенной функции положительный, то некоторая конечная разность становится уже константой (в приведенном примере третья), а следующие конечные разности тождественно равны нулю.

Аналогичным свойством последовательного понижения степени обладают ***конечные разности полиномов***. Пусть полином степени n представлен в стандартной форме:

$$P_n(x) = \sum_{i=0}^{n} c_i \cdot x^i \,, \tag{1.4}$$

где c_i ($i = 0, 1, \ldots, n$) – известные числовые коэффициенты.

Пользуясь определением (1.1), запишем первую конечную разность этого полинома:

$$\Delta P_n(x) = P_n(x+h) - P_n(x) = \sum_{i=0}^{n} c_i \,[(x+h)^i - x^i]. \tag{1.5}$$

Круглые скобки раскрываем по биному Ньютона. В каждом i-м слагаемом суммы (1.5) взаимно уничтожаются степени x^i со старшим показателем i. В последнем, n-м слагаемом суммы (1.5) уничтожится член со степенью x^n, а в остальных слагаемых он отсутствует. Следовательно, первая конечная разность $\Delta P_n(x)$ есть полином степени $(n-1)$.

Соберем члены с одинаковыми степенями аргумента x в сумме (1.5) и перепишем конечную разность (1.5) в виде:

$$\Delta P_n(x) = \sum_{i=0}^{n-1} b_i \cdot x^i \,, \tag{1.6}$$

где b_i ($i = 0, 1, \ldots, n-1$) – определенные коэффициенты, которые выражаются через величины c_i и шаг h.

Для дальнейшего нам потребуется коэффициент b_{n-1} при старшей степени аргумента x^{n-1}. Преобразования суммы (1.5) к виду (1.6) дают, что b_{n-1} равен nhc_n.

Теперь найдем вторую конечную разность исходного полинома $P_n(x)$. Согласно определению (1.2) нетрудно получить, что вторая конечная разность $\Delta^2 P_n(x)$ есть полином степени $(n-2)$, который можно представить в виде следующей суммы с коэффициентами d_i ($i = 0, 1, \ldots, n-2$):

$$\Delta^2 P_n(x) = \sum_{i=0}^{n-2} d_i \cdot x^i. \tag{1.7}$$

Расчет коэффициента d_{n-2} при старшем члене полинома (1.7) дает значение:

$$d_{n-2} = (n-1)\, h\, b_{n-1} = n(n-1)\, h^2\, c_n. \tag{1.8}$$

Продолжая рассчитывать конечные разности более высоких порядков, получим, что n-я конечная разность исходного полинома (1.4) равна постоянной величине:

$$\Delta^n P_n(x) = n!\, h^n\, c_n. \tag{1.9}$$

Очевидно, что следующие конечные разности тождественно равны нулю: $\Delta^k P_n(x) = 0$, где $k > n$.

Вычисления конечных разностей можно рассматривать как действие некоторого оператора Δ на функцию $f(x)$, которое определяется соотношением (1.1):

$$\Delta\,[f(x)] = f(x+h) - f(x).$$

Непосредственно из определения доказывается свойство линейности оператора Δ:

$$\Delta(u + v) = \Delta u + \Delta v\,, \quad \Delta(Cy) = C\Delta y, \tag{1.10}$$

где u, v и y – произвольные функции аргумента x, а C – константа.

Также из определения оператора Δ следует полезное свойство конечных разностей:

$$\Delta^m(\Delta^n y) = \Delta^{m+n} y\,. \tag{1.11}$$

Для общности последующих выражений доопределим понятие конечной разности нулевого порядка:

$$\Delta^0 y = y. \tag{1.12}$$

Теперь, пользуясь определением оператора Δ, можно записать:

$$f(x+h) = f(x) + \Delta f(x) \qquad \text{или} \qquad f(x+h) = (1+\Delta) f(x),$$

где символ Δ используется как оператор.

Использование последней формулы n раз дает соотношение:

$$f(x+nh) = (1+\Delta)^n f(x)\,. \tag{1.13}$$

Степень круглой скобки можно выразить по биному Ньютона:

$$f(x+nh) = \sum_{k=0}^{n} C_n^k \cdot \Delta^k f(x)\,, \tag{1.14}$$

где C_n^k – биномиальный коэффициент (число сочетаний из n по k).

Последнее выражение означает, что значения произвольной функции в $(n+1)$ равноотстоящих точках выражаются через конечные разности порядков от 0 до n.

Наконец, найдем общее выражение для n-й конечной разности произвольной функции $f(x)$. Для этого воспользуемся очевидным операторным тождеством:

$$\Delta = (1+\Delta) - 1.$$

Возведем это тождество в n-ю степень, и подействуем им на произвольную функцию $f(x)$, причем в правой части используем бином Ньютона:

$$\Delta^n f(x) = [(1+\Delta) - 1]^n f(x) =$$

$$= (1+\Delta)^n \cdot f(x) - C_n^1\ (1+\Delta)^{n-1} f(x) + C_n^2\ (1+\Delta)^{n-2} f(x) - \ldots + (-1)^n f(x).$$

Используя уравнение (1.13), представим выражение n-й конечной разности функции $f(x)$ в виде следующей знакопеременной суммы:

$$\Delta^n f(x) = f(x+nh) - C_n^1 f[x+(n-1)h] + C_n^2 f[x+(n-2)h] - \ldots (-1)^n f(x).$$

(1.15)

То же самое выражение можно получить, n-кратно применяя исходные определения (1.1) и (1.2).

Следовательно, n-я конечная разность любой функции f(x) выражается через n последовательных значений этой функции, вычисляемых в равноотстоящих точках, интервалы между которыми задаются постоянным шагом h.

Часто в задачах, решаемых численными методами, некоторая функция $f(x)$ задается таблицей значений $y_i = f(x_i)$ в нескольких точках x_i (i = =1, ... , n, где n – определенное конечное число). Тогда конечные разности заданной функции в точках x_i удобно вычислять по определению (1.1) и рекуррентному соотношению (1.2) и при этом размещать в таблице специального вида, приведенной в следующем примере.

Пример 2. Функция f (x) задана в десяти равноотстоящих точках x_i (i= 1, ... , 10), которые указаны во втором столбце таблицы 1.1. В третьем столбце записаны соответствующие значения функции $y_i = f$ (x_i). В первом столбце приведены индексы значений (порядковые номера).

Таблица 1.1

i	x_i	y_i	Δy_i	$\Delta^2 y_i$	$\Delta^3 y_i$	$\Delta^4 y_i$
1	0,1	2,9850	-0,0448	-0,0294	0,0008	0,0001
2	0,2	2,9402	-0,0742	-0,0286	0,0009	0,0006
3	0,3	2,8660	-0,1028	-0,0277	0,0015	-0,0001
4	0,4	2,7632	-0,1305	-0,0262	0,0014	0,0005
5	0,5	2,6327	-0,1567	-0,0248	0,0019	0,0001
6	0,6	2,4760	-0,1815	-0,0229	0,0020	0,0003
7	0,7	2,2945	-0,2044	-0,0209	0,0023	
8	0,8	2,0901	-0,2253	-0,0186		
9	0,9	1,8648	-0,2439			
10	1,0	1,6209				

Остальные столбцы служат для размещения конечных разностей от первого до четвертого порядков, вычисленных в точках x_i.

Каждая конечная разность, расположенная в клетке таблицы 1.1, получена как разность значений в соседнем столбце слева, причем вычитаемым является число в той же строке, а уменьшаемым – число в

следующей. Пустой правый нижний угол таблицы объясняется тем, что исходных данных недостаточно для расчета соответствующих конечных разностей по определению (1.2). Согласно выражению (1.15), для вычисления n-й конечной разности функции $f(x)$ в некоторой точке x_0 требуются n значений этой функции: в точке x_0 и еще в $(n–1)$ равноотстоящих точках $x_i = x_0 + i\,h$ $(i = 1, \ldots, n - 1)$.

Конечные разности, определенные формулами (1.1) и (1.2) и рассмотренные в данном параграфе, строго математически называются правыми конечными разностями. Для решения большинства прикладных задач вполне достаточно таких конечных разностей. Справедливости ради заметим, что в математике определяются и используются также левые и центральные конечные разности. Для знакомства с ними отсылаем читателя к более полным курсам численных методов и к специальной литературе [1,3,4].

Глава 2. ИНТЕРПОЛЯЦИЯ

§ 2.1. Постановка проблемы интерполяции

Одной из важнейших проблем вычислительной математики является приближение функций. Часто при решении прикладных задач приходится сталкиваться со следующей ситуацией.

Из теоретических соображений следует, что некоторая величина y является функцией $f(x)$ непрерывного аргумента x, причем явный вид функции $f(x)$ неизвестен. В то же время известно конечное количество значений функции y_i на ограниченном интервале аргумента $x \in [a, b]$ в точках x_i ($i = 0, 1, \ldots, n$). Эти точки x_i ($i = 0, 1, \ldots, n$) называются узлами. Иначе говоря, исходная информация об исследуемой функции может быть записана в виде таблицы, содержащей $n + 1$ упорядоченных пар чисел.

Таблица 2.1

x_0	x_1	…	x_n
$y_0 = f(x_0)$	$y_1 = f(x_1)$	…	$y_n = f(x_n)$

Пример графического изображения исходных данных таблицы 2.1 приведен на рис. 2.1.

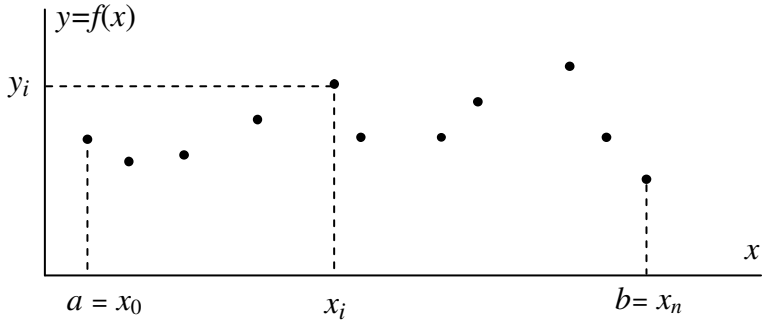


Рис. 2.1. Пример графика исходных данных, предназначенных для интерполяции. Черными кружками обозначены значения y_i функции $f(x)$ в узлах интерполяции x_i ($i = 0, 1, \ldots, n$).

Проблема заключается в вычислении значения функции $y = f(x)$ для произвольного аргумента x из интервала $[a, b]$, но не совпадающего ни с одним из узлов x_i ($i = 0, 1, \ldots, n$). Численные методы решения поставленной проблемы можно разделить на две группы, называемые ***интерполяцией*** и ***аппроксимацией.*** Методы интерполяции рассмотрены в данной главе, аппроксимации посвящена следующая глава.

Принцип интерполяции заключается в построении новой функции $F(x)$, которая называется ***интерполирующей функцией*** или ***интерполянтом***. Функция $F(x)$ принадлежит к определенному классу и в точках x_i ($i = 0, 1, \ldots, n$) принимает те же значения, что и функция $f(x)$

$$F(x_i) = y_i = f(x_i), \qquad i = 0, 1, \ldots, n. \tag{2.1}$$

Значения аргумента x_i ($i = 0, 1, \ldots, n$) при этом называются ***узлами интерполяции***.

Прежде чем переходить к определенным типам интерполянтов, сделаем два замечания. Во-первых, сами по себе данные таблицы 2.1 не могут определить конкретный вид интерполирующей функции. Например, точки графика на рис. 2.1 можно последовательно соединить отрезками прямых. С другой стороны, всегда можно найти алгебраический полином n-й степени с действительными коэффициентами, график которого точно проходит через $n + 1$ заданных точек, если все узлы таблицы 2.1 различны. Вообще существует бесконечное количество функций $F(x)$, удовлетворяющих условию (2.1). В качестве интерполирующих функций на практике часто используют алгебраические полиномы, суммы экспонент, Фурье-суммы, сплайны и т.д. На выбор типа интерполянта в конкретной задаче влияет любая дополнительная информация о зависимости между переменными x и y.

Второе замечание базируется на свойстве интерполянта (2.1) принимать в узлах точные значения интерполируемой функции. Если значения функции y_i таблицы 2.1 имеют существенную погрешность, то точное выполнение требования (2.1) является неосновательным. В таких ситуациях при вычислении значений функции $f(x)$ для аргументов x, не совпадающих с узлами, целесообразно применять методы аппроксимации, рассматриваемые в следующей главе.

§ 2.2. Интерполяционный полином Лагранжа

Алгебраические полиномы с действительными коэффициентами являются хорошо изученными функциями и весьма простыми для вычислений. Их удобно складывать и перемножать, дифференцировать и интегрировать. Алгебраический полином степени n можно записать в стандартной форме (1.4), хотя допустимы и другие представления. Из-за простоты работы с алгебраическими полиномами их часто используют в качестве интерполянтов.

Пусть исходной информацией являются $(n + 1)$ пар чисел x_i, y_i ($i = 0, 1, ..., n$), сведенных в таблицу вида 2.1. Будем строить интерполирующую функцию в виде алгебраического полинома.

Из теории функций известно, что через $n + 1$ произвольных точек на плоскости, заданных координатами x_i, y_i ($i = 0, 1, ... , n$), всегда можно провести график алгебраического полинома с действительными коэффициентами, если числа x_i ($i = 0, 1, ... , n$) различны. Если степень полинома $m = n$, то, вообще говоря, такой полином является единственным. Согласно требованию (2.1), в узлах интерполяции x_i ($i = 0, 1, ... , n$) значения этого полинома $P_m(x)$ точно совпадают с данными значениями y_i ($i= 0, 1, ... , n$) интерполируемой функции $f(x)$:

$$P_m(x_i) = y_i \qquad (i = 0, 1, ... , n). \tag{2.2}$$

Таким образом, всегда можно найти полином $P_m(x)$ степени $m = n$, который может служить интерполирующей функцией.

В специальных случаях, когда точки x_i , y_i ($i = 0, 1, ..., n$) расположены особым образом, степень полинома может быть меньше n. Например, если данные точки расположены точно на одной прямой, то через них пройдет график линейной функции (полинома первой степени) независимо от количества данных точек.

При $m > n$ полиномов, графики которых проходят через заданные точки, бесконечно много.

Рассмотрим ситуацию, когда расстояния между узлами неодинаковы. В таком случае целесообразно построить ***интерполяционный полином Лагранжа***. Этот полином строится в виде суммы:

$$L_n(x) = \sum_{i=0}^{n} \ell_i(x) \cdot y_i \,, \qquad (2.3)$$

где $\ell_i(x)$ – функции аргумента x, называемые ***коэффициентами Лагранжа*** ($i = 0, 1, ..., n$).

По свойству интерполирующего полинома (2.1) должно выполняться условие:

$$L_n(x_i) = y_i \qquad (i = 0, 1, ..., n). \qquad (2.4)$$

Для выполнения (2.4) на коэффициенты Лагранжа налагается следующее требование:

$$\ell_i(x_j) = \delta_{ij} \quad (i, j = 0, 1, ..., n), \qquad (2.5)$$

где δ_{ij} – символ Кронекера ($\delta_{ij} = 0$ при $i \neq j$ и $\delta_{ij} = 1$ при $i = j$).

Чтобы обеспечить выполнение равенств (2.5), коэффициенты Лагранжа $\ell_i(x)$ представим в виде произведений:

$$\ell_i(x) = c_i (x - x_0)(x - x_1)... (x - x_{i-1})(x - x_{i+1})... (x - x_n), \qquad (2.6)$$

где c_i – некоторые постоянные величины, пока неизвестные.

В выражении для i-го коэффициента Лагранжа отсутствует множитель $(x - x_i)$, поэтому каждое i-е произведение (2.6) содержит n скобок, и каждый коэффициент Лагранжа представляет собой полином степени n.

Для определения явного вида постоянных величин c_i можно воспользоваться равенством $\ell_i(x_i) = 1$ ($i = 0, 1, ..., n$), согласно требованию (2.5). Тогда из (2.6) сразу следует, что $1 / c_i = \prod_{j=0, j\neq i}^{n} (x_i - x_j)$.

Таким образом, коэффициенты Лагранжа представляются произведениями отношений разностей:

$$\ell_i(x) = \prod_{j=0, j\neq i}^{n} \frac{(x - x_j)}{(x_i - x_j)} \,. \qquad (2.7)$$

Легко убедиться непосредственной подстановкой, что для каждого коэффициента (2.7) выполняется условие (2.5).

Интерполяционный полином в форме Лагранжа получается подстановкой найденных коэффициентов (2.7) в выражение (2.3):

$$L_n(x) = \sum_{i=0}^{n} \left(y_i \cdot \prod_{j=0,\, j\neq i}^{n} \frac{(x - x_j)}{(x_i - x_j)} \right). \qquad (2.8)$$

Для большей наглядности запишем полином Лагранжа 3-й степени в развернутом виде:

$$L_3(x) = y_0 \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} + y_1 \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} +$$

$$+ y_2 \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} + y_3 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}.$$

Использование полинома Лагранжа для интерполяции иллюстрируется следующим примером.

Пример 1. Пусть значения функции заданы таблицей 2.2. Требуется построить по ним полином Лагранжа.

Таблица 2.2

i	0	1	2	3	4
x_i	–2	–1	1,2	3,5	5
y_i	–4	0	3,8	–2	7

Так как таблица 2.2 содержит 5 узлов, интерполяционный полином Лагранжа должен иметь 4-ю степень. Подставляя эти числовые значения таблицы в общую формулу (2.8), находим явный вид полинома Лагранжа для данной задачи:

$$L_4(x) = -4 \frac{(x+1)(x-1{,}2)(x-3{,}5)(x-5)}{(-2+1)(-2-1{,}2)(-2-3{,}5)(-2-5)} +$$

$$+ 3{,}8\ \frac{(x+2)(x+1)(x-3{,}5)(x-5)}{(1{,}2+2)(1{,}2+1)(1{,}2-3{,}5)(1{,}2-5)} -$$
$$-2\ \frac{(x+2)(x+1)(x-1{,}2)(x-5)}{(3{,}5+2)(3{,}5+1)(3{,}5-1{,}2)(3{,}5-5)} +$$
$$+ 7\ \frac{(x+2)(x+1)(x-1{,}2)(x-3{,}5)}{(5+2)(5+1)(5-1{,}2)(5-3{,}5)}.$$

Сумма содержит только 4 слагаемых, так как коэффициент Лагранжа $\ell_1(x)$ умножается на нулевое значение $y_1 = 0$. Приведение полинома к стандартному виду даст следующее выражение:

$$L_4(x) = 0{,}0819538\ x^4 - 0{,}181866\ x^3 - 1{,}43424\ x^2 + 2{,}19964\ x + 3{,}37007.$$

График построенного полинома приведен на рис. 2.2.

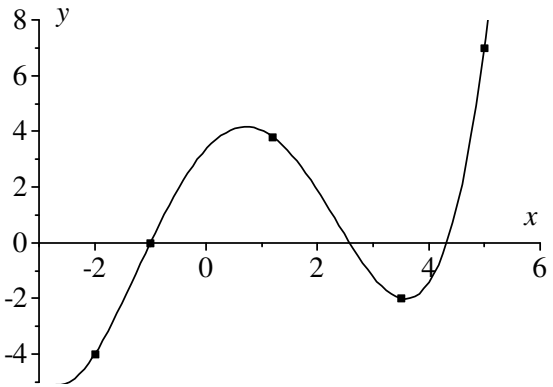


Рис. 2.2. Интерполяционный полином Лагранжа,
проведенный через 5 заданных точек таблицы 2.2.

Как было указано в начале данного параграфа, полином Лагранжа строится для произвольного набора узлов. Но вычисление значения полинома Лагранжа для заданного аргумента x по формуле (2.8) требует большого количества вычислений. Можно подсчитать, что при этом

необходимо выполнить $2n\,(n+1)+n$ сложений и вычитаний, а также $2n\,(n+1)$ умножений и делений.

В подробных курсах численных методов доказывается, что полином Лагранжа (2.8) обладает свойством единственности [1].

Полиномы, интерполирующие таблицы данных с постоянным шагом, позволяют проводить более быстрые вычисления, сравнительно с использованием универсального полинома Лагранжа.

§ 2.3. Интерполяция по равноотстоящим узлам

Рассмотрим случай, когда известны значения интерполируемой функции $f(x)$ в равноотстоящих узлах. При этом узлы интерполяции x_i выражаются формулой:

$$x_i = x_0 + i \cdot h, \qquad \text{где } i = 0, 1, \ldots, n\,. \tag{2.9}$$

Постоянный параметр h называется ***шагом интерполяции***.

В такой задаче исходными данными являются $(n+3)$ чисел: начальный узел интерполяции x_0, шаг интерполяции h и $(n+1)$ значений неизвестной функции y_i $(i = 0, 1, \ldots, n)$ в узлах интерполяции.

Будем строить полином $P_n(x)$ степени n, обладающий свойством (2.2)

$$P_n(x_i) = y_i \qquad (i = 0, 1, \ldots, n), \tag{2.10}$$

где значения x_i заданы формулой (2.9).

Полиномы для интерполяции по равноотстоящим узлам впервые были построены Ньютоном, который был как гениальным физиком, так и гениальным математиком.

Первый полином Ньютона строится в форме:

$$P_n(x) = a_0 + \sum_{i=1}^{n}\left(a_i \cdot \prod_{j=0}^{i-1}(x - x_j) \right) \tag{2.11}$$

или в развернутом виде:

$$P_n(x) = a_0 + a_1\,(x - x_0) + a_2\,(x - x_0)\,(x - x_1) + a_3\,(x - x_0)\,(x - x_1)\,(x - x_2) + \ldots + a_n\,(x - x_0)\ldots(x - x_{n-1}).$$

Коэффициенты полинома a_i необходимо выразить через известные величины x_0, h, n, y_i ($i = 0, 1, \ldots n$), пользуясь требованием (2.10), т.е. прохождением графика полинома через все точки заданной системы.

Сразу заметим, что подстановка $x = x_0$ в выражение (2.11) дает $P_n(x_0)=a_0$. Следовательно, согласно (2.10), мы получаем значение свободного члена искомого полинома:

$$a_0 = y_0. \tag{2.12}$$

Остальные коэффициенты полинома Ньютона выразим через конечные разности, описанные в предыдущей главе. Можно доказать эквивалентность условия (2.10) тому, что

$$\Delta^k P_n(x_0) = \Delta^k y_0 \qquad (k = 0, 1, \ldots, n). \tag{2.13}$$

Доказательство основано на том, что согласно (1.15), n-я конечная разность любой функции $f(x)$ выражается через n последовательных значений этой функции в равноотстоящих точках, отличающихся на шаг h.

Будем строить последовательно конечные разности $\Delta^i P_n(x)$, где ($i = 0, 1, \ldots, n$), для произвольной точки интерполяции x и заданного шага h.

Для нахождения первого коэффициента a_1 составим первую конечную разность полинома (2.11):

$$\Delta^1 P_n(x) = P_n(x+ h) - P_n(x). \tag{2.14}$$

Пользуясь свойством линейности конечных разностей, будем их вычислять для отдельных слагаемых суммы (2.11) и выносить за скобки общие множители a_i ($i = 0, 1, \ldots, n$):

$i = 0$ $\quad a_0 - a_0 = 0$

$i = 1$ $\quad a_1 [(x + h - x_0) - (x - x_0)] = a_1 h$

$i = 2$ $\quad a_2 [(x+ h - x_0) (x + h - x_1) - (x - x_0) (x - x_1)] =$

$= a_2 [(x + h - x_0) (x - x_0) - (x - x_0) (x - x_0 - h)] = a_2\, 2h(x - x_0).$

В последнем преобразовании использовалось, что $x_1 = x_0 + h$.

Дальнейшими вычислениями можно показать, что среди сомножителей остальных коэффициентов a_i ($i > 2$) обязательно

присутствуют множители $(x - x_0)$. В результате получим следующее для первой конечной разности исходного полинома:

$$\Delta^1 P_n(x) = a_1 h + 2a_2 (x - x_0) h + 3 a_3 (x - x_0) (x - x_1) h + ... \\ + n a_n (x - x_0)...(x - x_{n-2}) h. \qquad (2.15)$$

Теперь если в (2.15) положить $x = x_0$, получим: $\Delta^1 P_n(x_0) = a_1 h$. С другой стороны, из (2.13) и (2.10) следует: $\Delta^1 P_n(x_0) = y_1 - y_0$. Сравнивая два выражения для $\Delta^1 P_n(x_0)$, получаем представление первого коэффициента искомого полинома через известные величины:

$$a_1 = (y_1 - y_0) / h. \qquad (2.16)$$

Для вычисления коэффициента a_2 рассматривается вторая конечная разность полинома (2.11):

$$\Delta^2 P_n(x) = \Delta^1 P_n(x + h) - \Delta^1 P_n(x).$$

Ясно, что она не будет содержать слагаемого с коэффициентом a_1, т.к. этот коэффициент является множителем константы в сумме $\Delta^1 P_n(x)$, согласно (2.15). Если провести преобразования, аналогичные проделанным для первой конечной разности исходного полинома, то мы получим, что свободным членом нового полинома $\Delta^2 P_n(x)$ будет $2a_2 h^2$, а остальные члены будут содержать сомножители $(x - x_0)$. Тогда ясно, что в точке $x = x_0$ получим:

$$\Delta^2 P_n(x_0) = 2a_2 h^2.$$

С другой стороны, по формуле (1.3) для произвольной функции имеем:

$$\Delta^2 P_n(x_0) = P_n(x_2) - 2P_n(x_1) + P_n(x_0),$$

а по условию (2.10): $\Delta^2 P_n(x_0) = y_2 - 2 y_1 - y_0$.

Сравнение дает выражение для второго коэффициента искомого полинома:

$$a_2 = (y_2 - 2 y_1 - y_0) / (2h^2). \qquad (2.17)$$

Для получения остальных коэффициентов искомого полинома (2.11) необходимо повторять аналогичную процедуру с высшими конечными разностями. В результате мы получим следующее общее выражение для коэффициентов интерполянта Ньютона:

$$a_i = \frac{\Delta^i(y_0)}{i!h^i}. \tag{2.18}$$

Следовательно, интерполяционный полином Ньютона выразится через конечные разности порядков 1, ... , n, вычисленные в нулевом узле x_0:

$$P_n(x) = y_0 + \sum_{i=0}^{n}\left[\frac{\Delta^i(y_0)}{i!h^i}\prod_{k=0}^{i-1}(x-x_k)\right]. \tag{2.19}$$

То, что построенный полином удовлетворяет условию (2.10), проверяется непосредственной подстановкой. Положим $x = x_j$, где x_j – произвольный узел, и подставим в (2.19). При этом учтем, что все разности $(x_j - x_k)$ кратны целому числу шагов h при $k \neq j$ и равны нулю при $k = j$. Следовательно, все слагаемые суммы (2.19) с индексами $i > j$ содержат множитель, равный нулю. В результате получим:

$$P_n(x_j) = y_0 + j\,\Delta y_0 + \frac{j\cdot(j-1)}{2}\,\Delta^2 y_0 + \ \dots\ + \frac{j\cdot(j-1)\cdot\ldots\cdot 1}{j!}\,\Delta^j y_0.$$

Использование (1.14) и (2.13) позволяет представить правую часть последнего уравнения в виде

$$P_n(x_0 + jh) = P_n(x_j) = y_j,$$

что и требовалось доказать.

Для практических расчетов полином Ньютона (2.19) удобнее представлять, вводя новую переменную:

$$q = (x - x_0)/h. \tag{2.20}$$

Тогда интерполяционный полином Ньютона представится так:

$$P_n(x) = y_0 + q\cdot\Delta^1 y_0 + \frac{q(q-1)}{2!}\cdot\Delta^2 y_0 + \frac{q(q-1)(q-2)}{3!}\cdot\Delta^3 y_0 + \ldots$$
$$+ \frac{q(q-1)\ldots(q-n+1)}{n!}\cdot\Delta^n y_0. \qquad (2.21)$$

Анализ формулы (2.21) показывает, что погрешность построенного интерполянта Ньютона принимает наименьшие значения для аргументов, близких к нулевому узлу x_0 (см. § 2.5). Пример, приведенный в предыдущем параграфе для полинома Лагранжа, показывает, что не всегда целесообразно строить полином по всем узлам, так при этом могут возникать необоснованные «выбросы». В практических расчетах часто заранее выбирается степень интерполирующего полинома. Затем в качестве нулевого x_0 берется узел, ближайший слева к значению аргумента x, для которого требуется вычислить значение неизвестной функции $f(x)$. Интерполянт n-й степени (2.21) строится с помощью значений функции y_i в узле x_0 и в n узлах, расположенных правее аргумента x. Значения функции в узлах левее x_0 не используются.

Пример 2. Исходные данные сведены в таблицу 2.3 с постоянным шагом $h = 0{,}1$.

Таблица 2.3

x_i	3,0	3,1	3,2	3,3	3,4	3,5	3,6
y_i	0,550	0,450	0,365	0,293	0,234	0,185	0,146

Пусть требуется вычислить значение функции $f(x)$ для аргумента x=3,22. Узлы в таблице 2.3 являются равноотстоящими (шаг h = 0,1), поэтому целесообразно построить интерполянт Ньютона.

Ограничимся полиномом 3-го порядка. За нулевой узел, как было рекомендовано ранее, возьмем ближайший слева к значению x = 3,22 , т.е. узел x_0 = 3,2. Полином будем строить по четырем узлам $x_i \geq 3{,}2$. Составим таблицу конечных разностей $\Delta^k y_i$ (k =1, 2, 3), необходимых для построения полинома (2.21) при n = 3. Таблицу построим по типу примера 2 предыдущей главы (см. табл. 1.1). Напомним, что разности порядка выше первого вычисляются как разности разностей предыдущего порядка, стоящие в соседнем левом столбце таблицы 2.4.

Таблица 2.4

i	x_i	y_i	Δy_i	$\Delta^2 y_i$	$\Delta^3 y_i$
0	3,2	0,365	– 0,072	0,013	– 0,003
1	3,3	0,293	– 0,059	0,01	
2	3,4	0,234	– 0,049		
3	3,5	0,185			

Для значений $x = 3{,}22$ и $x_0 = 3{,}2$ вычислим параметр $q = (3{,}22–3{,}2)/0{,}1 = 0{,}2$ и подставим все требуемые числа в формулу (2.21):

$$f(3{,}22) = P_3(x = 3{,}22) = 0{,}365 + 0{,}2\,(-0{,}072) + \frac{0{,}2(0{,}2-1)}{2}0{,}013 + \\ + \frac{0{,}2(0{,}2-1)(0{,}2-2)}{3!}(-0{,}003) \approx 0{,}349.$$

Исходные значения функции в узлах интерполирования были заданы с тремя значащими цифрами (см. табл.2.3), поэтому, согласно правилам, результат округляем до трех значащих цифр.

Приведем формулы для двух важных частных случаев интерполяции.

1. Линейная интерполяция. Интерполяционный полином (2.21) строится по двум узлам и является линейной функцией:

$$P_1(x) = y_0 + q\cdot\Delta^1 y_0 \quad \text{или} \quad P_1(x) = y_0 + \frac{x - x_0}{h}(y_1 - y_0). \quad (2.22)$$

Последнюю формулу легко получить с помощью рисунка 2.1, соединив точки $(x_0,\ y_0)$ и $(x_1,\ y_1)$ отрезком прямой линии. Соединяя отрезками прямых точки $(x_i,\ y_i)$ и $(x_{i+1},\ y_{i+1})$ для соседних узлов, можно получить график кусочно-линейной интерполяции. Практика показывает, что для большинства физических задач такой вид интерполяции неприменим, так как дает большую погрешность.

2. Квадратичная интерполяция. Для построения интерполяционного полинома (2.21) выбираются три узла. Полином становится квадратичной функций следующего вида:

$$P_2(x) = y_0 + q \cdot \Delta^1 y_0 + \frac{q(q-1)}{2} \cdot \Delta^2 y_0$$

или

$$P_2(x) = y_0 + q \cdot (y_1 - y_0) + \frac{q(q-1)}{2} \cdot (y_2 - 2\,y_1 + y_0). \qquad (2.23)$$

Выше было показано, что интерполянт (2.21) строится с использованием n значений функции в узлах данной таблицы, которые расположены правее аргумента x, для которого проводится интерполяция. Если же аргумент x находится ближе к концу таблицы, например, лежит в интервале (x_{n-1}, x_n), то построить полином вида (2.21) невозможно (за исключением линейного). В подобных случаях следует вместо полинома (2.21) использовать ***второй интерполяционный полином Ньютона*** (или полином для интерполирования назад). Этот интерполянт задается тоже в виде (2.11), но его коэффициенты выражаются через значения функции в узлах, расположенных левее заданного аргумента x. Например, если $(x_{n-1} < x < x_n)$, то для вычисления коэффициентов второго полинома Ньютона степени n требуются значения y_i $(i = 0, 1, ..., n)$. Расчет коэффициентов второго полинома Ньютона делается аналогично расчету для первого полинома. Подробный вывод приводится во всех курсах вычислительной математики (например, [1]), поэтому здесь сразу приведем вид второго интерполянта Ньютона в явной форме:

$$Q_n(x) = y_n + \frac{x - x_n}{h} \Delta y_{n-1} + \frac{(x - x_n)(x - x_{n-1})}{2h^2} \Delta^2 y_{n-2} + \\ + \frac{(x - x_n)...(x - x_1)}{n! h^n} \Delta^n y_0. \qquad (2.24)$$

Заметим, что коэффициенты второго интерполянта Ньютона представляются через конечные разности, вычисляемые в различных узлах. Напротив, все конечные разности, необходимые для получения первого интерполянта Ньютона, рассчитываются для одного узла, который выбран в качестве нулевого x_0.

Для облегчения практического использования интерполянта (2.24) целесообразно сделать следующую замену переменной

$$q = (x - x_n) / h. \qquad (2.25)$$

Тогда второй интерполяционный полином Ньютона приобретет вид:

$$Q_n(x) = y_n + q\,\Delta y_{n-1} + \frac{q(q+1)}{2}\Delta^2 y_{n-2} + \frac{q(q+1)(q+2)}{3!}\Delta^3 y_{n-3} + \\ + \frac{q(q+1)...(q+n-1)}{n!}\Delta^n y_0. \qquad (2.26)$$

Этот полином n-й степени строится по n точкам, расположенным левее аргумента x, и единственной точке x_n, лежащей правее x. Поэтому на практике второй полином Ньютона (2.26) используется, когда требуется вычислить значение неизвестной функции $f(x)$ для аргумента x, находящегося в конце заданной таблицы узлов.

Если для интерполяции используется полином степени $m < n - 1$ (где n – количество узлов таблицы), то в качестве узла x_n выбирается ближайший справа к аргументу x, для которого проводится вычисление интерполянта. Полином Ньютона (2.26) строится по значению функции y_n и значениям функции в m узлах, расположенных левее аргумента x. При этом величина (2.25) меньше единицы и погрешность интерполяции с помощью второго полинома Ньютона минимальна (см. § 2.5).

Пример 3. Вычислим значение функции $f(x)$, заданной таблицей 2.3, в точке $x = 3{,}43$. На основе общей формулы (2.26) построим второй интерполяционный полином Ньютона 3-го порядка, причем за узел x_n примем значение $x_3 = 3{,}5$:

$$Q_3(x) = y_3 + q\,\Delta y_2 + \frac{q(q+1)}{2}\Delta^2 y_1 + \frac{q(q+1)(q+2)}{3!}\Delta^3 y_0.$$

Значения конечных разностей, необходимые для расчета, возьмем из таблицы 2.4. Параметр q, согласно определению (2.25), равен $(3{,}43-3{,}5)/0{,}1=-0{,}7$. Расчет значения $f(x=3{,}43)$ по последней формуле дает

$$f(3{,}43) = Q_3(x=3{,}43) = \\ =0{,}185+0{,}7\cdot 0{,}049-0{,}7\cdot 0{,}3\cdot 0{,}01/2+0{,}7\cdot 0{,}3\cdot 1{,}3\cdot 0{,}003/6 \approx 0{,}218.$$

Если удается построить таблицу значений интерполируемой функции для равноотстоящих узлов вида (2.9), то предпочтительнее использовать полиномы Ньютона, а не полином Лагранжа. Для вычисления одного значения полинома (2.21) или (2.26) при рациональном проведении расчетов требуется выполнить $n\,(n+1)/2+n$ сложений и вычитаний плюс $3n$ умножений и делений. Это значительно меньше, чем при вычислении полинома Лагранжа того же порядка.

Оба интерполянта Ньютона используют значения функции в узлах, лежащих по одну сторону от выбранного значения аргумента x.

В вычислительной математике также применяются интерполяционные полиномы, построенные на системе равноотстоящих узлов, коэффициенты которых рассчитываются по значениям функции в узлах, расположенных как левее, так и правее заданной точки x.

Для построения таких интерполянтов значения узлов удобнее задавать в следующей форме

$$x_i = x_0 + i\cdot h\,, \quad \text{где } i = 0, \pm1, \pm2, \ldots \pm n. \tag{2.27}$$

Таким образом, номера узлов принимают как положительные, так и отрицательные значения. Центральный узел имеет нулевой индекс.

Будем далее в этом параграфе использовать величину q, определенную формулой (2.20). Приведем без вывода интерполяционную формулу Стирлинга в следующем виде:

$$\begin{aligned} P_S(x) = y_0 + q\frac{\Delta y_{-1}+\Delta y_0}{2} + \frac{q^2}{2}\Delta^2 y_{-1} + \frac{q(q^2-1)}{3!}\,\frac{\Delta^3 y_{-2}+\Delta^3 y_{-1}}{2} + \\ + \frac{q^2(q^2-1)}{4!}\Delta^4 y_{-2} + \frac{q(q^2-1)(q^2-4)}{5!}\,\frac{\Delta^5 y_{-3}+\Delta^5 y_{-2}}{2} + \\ + \frac{q(q^2-1)(q^2-4)}{6!}\Delta^6 y_{-3} + \ldots + \frac{q(q^2-1)(q^2-4)(q^2-9)\ldots(q^2-(n-1)^2)}{(2n-1)!}\times \\ \times\frac{\Delta^{2n-1}y_{-n}+\Delta^{2n-1}y_{-(n-1)}}{2} + \\ + \frac{q^2(q^2-1)(q^2-4)(q^2-9)\ldots(q^2-(n-1)^2)}{(2n)!}\cdot\Delta^{2n}y_{-n}. \end{aligned} \tag{2.28}$$

В последнем выражении символами $\Delta^k y_i$ обозначены конечные разности, определенные формулами:

$$\Delta y_i = y_{i+1} - y_i \; ; \; \Delta^2 y_i = \Delta y_{i+1} - \Delta y_i \;\text{ и т.д.} \qquad (2.29)$$

Специальные исследования показали, что интерполяционная формула Стирлинга (2.28) имеет наименьшую погрешность, когда абсолютная величина $|q| \le 1/4$.

Если $1/4 \le q \le 3/4$, то более лучшую точность дает интерполяционная формула Бесселя:

$$P_B(x) = \frac{y_0 + y_1}{2} + (q - 1/2)\,\Delta y_0 + \frac{q(q-1)}{2} \cdot \frac{\Delta^2 y_{-1} + \Delta^2 y_0}{2} +$$

$$+ \frac{(q-1/2)q(q-1)}{3!}\Delta^3 y_{-1} + \frac{q(q-1)(q+1)(q-2)}{4!} \cdot \frac{\Delta^4 y_{-2} + \Delta^4 y_{-1}}{2} +$$

$$+ \frac{q(q-1)(q+1)(q-2)(q+2)...(q-n)(q+n-1)}{(2n)!} \cdot \frac{\Delta^{2n} y_{-n} + \Delta^{2n} y_{-n+1}}{2} +$$

$$+ \frac{(q-1/2)q(q-1)(q+1)(q-2)(q+2)...(q-n)(q+n-1)}{(2n+1)!}\Delta^{2n+1} y_{-n}. \qquad (2.30)$$

В качестве частных случаев приведем квадратичные интерполяционные формулы Стирлинга:

$$P_S(x) = y_0 + q\frac{\Delta y_{-1} + \Delta y_0}{2} + \frac{q^2}{2}\Delta^2 y_{-1} =$$

$$= y_0 + \frac{q}{2}\cdot(y_0 - y_{-1}) + \frac{q^2}{2}\,(y_1 + y_{-1} - 2\,y_0) \qquad (2.31)$$

и Бесселя:

$$P_B(x) = \frac{y_0 + y_1}{2} + (q - 1/2)\,\Delta y_0 + \frac{q(q-1)}{2} \cdot \frac{\Delta^2 y_{-1} + \Delta^2 y_0}{2} =$$

$$= y_0 + q \cdot (y_1 - y_0) + \frac{q(q-1)}{4}\,(y_2 - y_1 - y_0 + y_{-1}). \qquad (2.32)$$

Видно, что для вычисления интерполяционного полинома Стирлинга 2-го порядка надо знать значения интерполируемой функции в трех равноотстоящих узлах. Для вычисления интерполяционного полинома Бесселя 2-го порядка необходимо иметь значения функции уже в четырех узлах.

§ 2.4. Сплайн - интерполяция

Помимо «классических» интерполяционных полиномов, в XX веке получила широкое распространение в практике еще одна разновидность интерполянтов, называемых ***сплайнами***. Исторически они возникли в авиастроении при конструировании обтекаемых профилей.

Сплайном $S(x)$ является интерполирующая функция, которая обладает наименьшей кривизной и имеет квадратично интегрируемую вторую производную. Так как сплайн является интерполянтом, то для него должно выполняться условие (2.1):

$$S(x_i) = y_i = f(x_i) \qquad (i = 0, 1, \ldots, n). \qquad (2.33)$$

Вид функции сплайна $S(x)$ находится из условия минимальности функционала

$$\int_{x_0}^{x_n} \left(S''(x)\right)^2 dx\,. \qquad (2.34)$$

Эта функция физически моделируется гибким и упругим стержнем, который проходит через все заданные точки интерполяции. Минимум функционала (2.34) соответствует минимуму упругой потенциальной энергии этого стержня.

Особенно часто используется ***естественный кубический сплайн*** – полином 3-й степени, удовлетворяющий требованиям (2.33) и (2.34). Такой естественный кубический сплайн является непрерывной

функцией и имеет непрерывные первую и вторую производную на интервале интерполяции (x_0 , x_n).

Как и в предыдущих параграфах данной главы, интерполянт строится на базе таблицы данных типа (2.9).

Таблица с $n + 1$ узлами определяет n подинтервалов между этими узлами. Кубический полином (естественный сплайн) необходимо построить для каждого подинтервала. Таким образом, интерполяция кубическими сплайнами относится к классу кусочно-многочленной интерполяции. Так как каждый кубический полином характеризуется четырьмя константами, поэтому для построения сплайна $S(x)$ на всем интервале $[x_0; x_n]$ необходимо определить $4n$ параметров.

Требование непрерывности функции, ее первой и второй производных во внутренних $(n - 1)$ узлах дает $3(n - 1)$ уравнения для определения неизвестных параметров. Условие (2.33) для всех $(n + 1)$ узлов дает еще $(n + 1)$ уравнений. Недостающие два условия для однозначного определения естественного сплайна задаются приравниванием нулю вторых производных в точках x_0 и x_n

$$S''(x_0) = S''(x_n) = 0, \tag{2.35}$$

что означает нулевую кривизну сплайна на концах интервала.

Иногда при построении сплайнов задаются другие граничные условия на концах интервала интерполяции для самих функций или первых производных.

После формулировки общих положений займемся конструированием явного вида коэффициентов кубического сплайна. Для этого вначале рассмотрим отдельный подинтервал (x_i , x_{i+1}). Введем обозначения

$$h_i = (x_{i+1} - x_i), \tag{2.36}$$

$$w = (x - x_i) / h_i, \tag{2.37}$$

$$v = 1 - w = (x_{i+1} - x) / h_i. \tag{2.38}$$

Сплайн на подинтервале (x_i , x_{i+1}) удобно представить формулой:

$$S(x) = w\, y_{i+1} + v\, y_i + h_i^2[\, (w^3 - w)\, \sigma_{i+1} + (v^3 - v)\sigma_i\,], \tag{2.39}$$

где σ_{i+1} и σ_i – константы, которые необходимо выразить через известные величины x_i, y_i. Очевидно, что функция (2.39) – полином 3-й степени относительно аргумента x.

Легко убедиться непосредственной подстановкой, что

$$S(x_i) = y_i, \quad S(x_{i+1}) = y_{i+1}. \tag{2.40}$$

Получим выражения для трех производных функции (2.39):

$$S'(x) = (y_{i+1} - y_i)/ h_i + h_i[(3w^2 - 1) \sigma_{i+1} - (3v^2 - 1)\sigma_i], \tag{2.41}$$

$$S''(x) = 6 (w\sigma_{i+1} + v\sigma_i), \tag{2.42}$$

$$S'''(x) = 6 (\sigma_{i+1} - \sigma_i) / h_i . \tag{2.43}$$

Формула (2.41) позволяет получить выражение для первой производной на концах любого подинтервала. Декларированная непрерывность первой производной сплайна $S(x)$ на всем интервале (x_0 , x_n) позволяет приравнять правые и левые производные в точках узлов x_i. В результате получается соотношение

$$h_{i-1} \sigma_{i-1} + 2(h_{i-1} + h_i) \sigma_i + h_i \sigma_{i+1} = \Delta_i - \Delta_{i-1}. \tag{2.44}$$

В уравнении (2.44) и последующих этого параграфа использовано новое обозначение:

$$\Delta_i = (y_{i+1} - y_i) / h_i . \tag{2.45}$$

Заметим, что в этом параграфе величины Δ_i, определенные формулой (2.45), называются разделенными разностями, в отличие от конечных разностей (1.1), введенных в главе 1.

Соотношения (2.44) получены для всех n – 1 внутренних узлов. Индекс i пробегает значения 1, ..., n–1.

Таким образом, получена система (n – 1) линейных уравнений относительно неизвестных величин σ_i.

В качестве недостающих условий можно взять уравнения, выражающие требования (2.35). Можно поступить иначе – использовать единственные кубические полиномы, которые проходят точно через 4 первые и 4 последние точки из заданных x_i, y_i (i = 0,. . ., n). Тогда два

дополнительных условия на концах интервала (x_0 , x_n) выразятся следующим образом:

$$(\sigma_1 - \sigma_0) / h_0 = \Delta_0^{(3)}, \quad (\sigma_n - \sigma_{n-1}) / h_{n-1} = \Delta^{(3)}_{n-3}, \quad (2.46)$$

где

$$\Delta_i^{(3)} = (\Delta^{(2)}_{i+1} - \Delta^{(2)}_i) / (x_{i+3} - x_i), \Delta_i^{(2)} = (\Delta_{i+1} - \Delta_i) / (x_{i+2} - x_i). \quad (2.47)$$

Полученная система линейных уравнений (2.44) и (2.46) для нахождения $n + 1$ параметров σ_i ($i = 0,. . ., n$) имеет единственное решение. Матрица коэффициентов при неизвестных σ_i может быть записана в следующем виде:

$$\begin{bmatrix} -h_0 & h_0 & 0 & 0 & 0 \\ h_0 & 2(h_0 + h_1) & h_1 & 0 & 0 \\ 0 & h_1 & 2(h_1 + h_2) & h_2 & 0 \\ . & . & . & . & . \\ 0 & 0 & h_{n-2} & 2(h_{n-2} + h_{n-1}) & h_{n-1} \\ 0 & 0 & 0 & h_{n-1} & -h_{n-1} \end{bmatrix}. \quad (2.48)$$

Правые части уравнений (свободные члены) соответственно равны:

$$h_0^2 \Delta_0^{(3)} , \Delta_1 - \Delta_0 , \Delta_2 - \Delta_1 , \ldots , \Delta_{n-1} - \Delta_{n-2} , -h_{n-1}^2 \Delta_{n-3}^{(3)} .$$

Решение полученной системы линейных уравнений удобно реализовать, пользуясь трехдиагональностью матрицы коэффициентов при неизвестных, применяя методы, приведенные в главе 4 настоящего пособия. В результате искомые величины σ_i представятся в виде рекуррентных выражений:

$$\sigma_n = \beta_n / \alpha_n , \quad \sigma_i = (\beta_i - h_i \sigma_{i+1}) / \alpha_i , ...i = n - 1, n - 2, \ldots, 0, \quad (2.49)$$

где коэффициенты α_i и β_i выражаются так:

$$\begin{aligned} &\alpha_0 = - h_0, \\ &\alpha_i = 2 (h_{i-1} + h_i) - h^2_{i-1} / \alpha_{i-1}, \quad i = 1, \ldots, n - 1 \\ &\alpha_n = - h_{n-1} - h^2_{n-1} / \alpha_{n-1} \end{aligned} \quad (2.50)$$

$$\beta_0 = h_0^2\,\Delta_0^{(3)}$$
$$\beta_i = (\Delta_i - \Delta_{i-1}) - h_{i-1}\,\beta_{i-1} / \alpha_{i-1}, \quad i = 1, \ldots, n-1 \qquad (2.51)$$
$$\beta_n = -h^2_{n-1}\Delta^{(3)}_{n-3} - h_{n-1}\,\beta_{n-1} / \alpha_{n-1}.$$

При необходимости многократно вычислять сплайн его удобнее представить в несколько иной форме:

$$S(x) = y_i + b_i\,(x - x_i) + c_i\,(x - x_i)^2 + d_i\,(x - x_i)^3 \qquad (2.52)$$

где $x_i \le x \le x_{i+1}, \quad i = 0, \ldots, n-1$.

Преобразование функции (2.39) к форме (2.52) дает следующие выражения для искомых коэффициентов:

$$b_i = (y_{i+1} - y_i) / h_i - h_i\,(\sigma_{i+1} + 2\,\sigma_i)\,, \quad c_i = 3\sigma_i\,, \quad d_i = (\sigma_{i+1} - \sigma_i) / h_i\,, \quad (2.53)$$

где $i = 0, \ldots, n-1$.

Подстановка коэффициентов (2.53) в выражение (2.52) дает функцию сплайна в явном виде для аргументов $x_i \le x \le x_{i+1}$, лежащих в i-м поддиапазоне интерполяции.

Пример 4. Пусть необходимо построить сплайн-интерполянт по набору шести точек, сведенных в таблицу 2.5.

Таблица 2.5

i	0	1	2	3	4	5
x_i	1	3	6	7	8	9
y_i	2	4	7	7	6	5

Для вычисления коэффициентов сплайна целесообразно по формулам (2.36) – (2.38), (2.43) и (2.45) вычислить промежуточные данные и свести их в таблицу 2.6.

Таблица 2.6

i	x_i	y_i	h_i	Δ_i	$\Delta_i^{(2)}$	$\Delta_i^{(3)}$
0	1	2	2	1	0	−1/24
1	3	4	3	1	−1/4	−1/20
2	6	7	1	0	−1/2	1/6

3	7	7	1	–1	0	
4	8	6	1	–1		
5	9	5				

Пользуясь величинами таблицы 2.6 можно составить систему линейных уравнений (2.44) и (2.46). Запишем эту систему в матричном виде:

$$\begin{bmatrix} -2 & 2 & 0 & 0 & 0 & 0 \\ 2 & 10 & 3 & 0 & 0 & 0 \\ 0 & 3 & 8 & 1 & 0 & 0 \\ 0 & 0 & 1 & 4 & 1 & 0 \\ 0 & 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \cdot \begin{pmatrix} \sigma_0 \\ \sigma_1 \\ \sigma_2 \\ \sigma_3 \\ \sigma_4 \\ \sigma_5 \end{pmatrix} = \begin{pmatrix} -1/6 \\ 0 \\ -1 \\ -1 \\ 0 \\ -1/6 \end{pmatrix}. \qquad (2.54)$$

Решением системы (2.54) является следующий набор значений

$$\sigma_i \ (i = 0, 1,. . ., 5) = \{67/708, 2/177, –107/1062, –121/531, 13/1062, 95/531\},$$

которые мы представили в виде обыкновенных дробей.

Вычисленные значения σ_i и h_i (i = 0, 1,. . ., 5) подставляются в формулы (2.36) – (2.39) или (2.52), (2.53), по которым можно получить явный вид сплайна $S(x)$ на каждом i-м поддиапазоне $x_i \le x \le x_{i+1}$.

Например, на интервале $x \in [3, 6]$ (т.е. между узлами i=1 и i=2) сплайн-интерполянт может быть выражен аналитически следующим кубическим полиномом:

$$S_{12}(x) = 1{,}6102 + 0{,}0226\,x + 0{,}3701\,x^2 - 0{,}0374\,x^3\,, \qquad (2.55)$$

где коэффициенты вычислены с точностью до 0,0001. График полученного сплайна представлен на рис. 2.3.

Аналогичным образом можно построить кубические интерполирующие полиномы на всех интервалах, ограниченных узлами таблицы 2.5. В результате получается сплайн-интерполянт для диапазона $1 \le x \le 9$, график которого изображен на рис. 2.4.

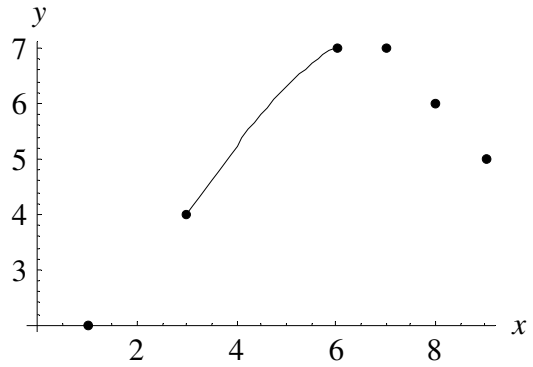


Рис. 2.3. График сплайн-интерполянта на интервале $x \in [3, 6]$. Кружки – исходные данные таблицы 2.4, линия – график полинома (2.55).

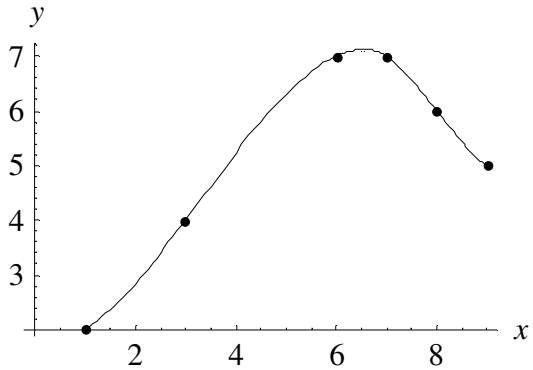


Рис. 2.4. График сплайн-интерполянта на интервале $x \in [1, 9]$. Кружки – исходные данные таблицы 2.4.

§ 2.5. Погрешность интерполяционных формул

Количественной характеристикой погрешности интерполяционной формулы является остаточный член, который определяется разностью:

$$R(x) = f(x) - P(x), \tag{2.56}$$

где $f(x)$ – интерполируемая функция, а $P(x)$ – интерполирующий полином. Абсолютная величина остаточного члена может быть принята за оценку погрешности вычисления значения функции $f(x)$ с помощью выбранного интерполирующего полинома.

Нам известны только значения функции $f(x)$ в конечном числе узлов, поэтому для аргументов x, не совпадающих с узлами, величину остаточного члена $R(x)$ можно оценить только приближенно.

Расчет значений остаточного члена $R(x)$ на интервале $[a, b]$ проводится обычным способом математического анализа (с помощью теоремы Ролля) и описан в подробных курсах численных методов [1, 3, 4]. Абсолютная величина остаточного члена для полинома Лагранжа n-й степени не превышает следующего значения

$$| R_L | \le \frac{M_{n+1}}{(n+1)!} \prod_{i=0}^{n} (x - x_i) \ , \tag{2.57}$$

где M_{n+1} обозначает максимум абсолютного значения производной $(n+1)$-го порядка функции $f(x)$ на интервале интерполирования $[a, b]$:

$$M_{n+1} = \max_{a \le x \le b} \left| f^{(n+1)}(x) \right| \ . \tag{2.58}$$

Границы интервала $[a, b]$, как и в предыдущих параграфах, совпадают с крайними узлами: $a = x_0$, $b = x_n$.

Задача интерполирования, как правило, на практике решается, когда исходная функция $f(x)$ задана лишь таблицей значений x_i, $y_i = f(x_i)$, где i=0, 1, ..., n (см. табл. 2.1). В этих случаях, к сожалению, невозможно вычислять значения производной $(n+1)$-го порядка этой функции.

Если интерполянт строится с использованием части таблицы данных, то количество узлов может быть достаточным для оценки величины $f^{(n+1)}(x)$ методом численного дифференцирования (см.

главу 9). Однако в этом случае возникает дополнительная трудность оценки точности численного значения производной (n+1)-го порядка.

Согласно выражению (2.57), на величину погрешности полинома Лагранжа, кроме значения производной (n+1)-го порядка, существенно влияет расположение узлов на интервале интерполяции [a, b]. Исследования показали, что уменьшения остаточного члена полинома Лагранжа можно добиться, если узлы интерполяции расположить по правилу:

$$x_i = \frac{a+b}{2} + \frac{b-a}{2} t_i \,, \quad (i = 0, 1, ..., n) \tag{2.59}$$

где t_i – нули полинома Чебышева (n+1)-го порядка, равные

$$t_i = -\cos \frac{2i+1}{2n+2} \pi \,, \qquad i = 0, 1, ..., n \,. \tag{2.60}$$

В этом случае модуль остаточного члена полинома Лагранжа не превышает величины

$$| R_L | \le \frac{M_{n+1}}{(n+1)!} 2 \left(\frac{b-a}{4} \right)^{n+1} . \tag{2.61}$$

Пользуясь выражением (2.57) и постоянным шагом h таблицы узлов, можно получить оценку остаточного члена для первого интерполяционного полинома Ньютона

$$R_1 \le h^{(n+1)} \frac{q(q-1)...(q-n)}{(n+1)!} M_{n+1}, \tag{2.62}$$

где величина q определяется формулой (2.20).

Приближенное значение производной (n+1)-го порядка функции $f(x)$ можно получить с помощью конечных разностей:

$$M_{n+1} \approx \frac{\Delta^{n+1} y_0}{h^{n+1}} \,. \tag{2.63}$$

Чем меньше величина шага h , тем точнее это приближенное равенство.

Подстановка последнего выражения в (2.62) дает приближенное значение погрешности первого интерполяционного полинома Ньютона

$$| R_1 | \approx \frac{q(q-1)...(q-n)}{(n+1)!} \Delta^{n+1} y_0. \qquad (2.64)$$

Следует заметить, что для вычисления конечной разности (n+1)-го порядка для значения $x = x_0$ необходимо знать значения функции $f(x)$ в (n+1) узлах, расположенных правее x_0 (см. главу 1 и формулу (1.15)).

Из формулы (2.64) видно, что если значение аргумента x лежит в интервале (x_0 , x_1), то величина $q < 1$ и погрешность интерполяции минимальна.

Аналогично можно получить оценку погрешности для второго интерполяционного полинома Ньютона

$$| R_2 | \approx \frac{q(q+1)...(q+n)}{(n+1)!} \Delta^{n+1} y_n, \qquad (2.65)$$

где величина q определяется уже формулой (2.25).

Ясно, что для вычисления этой оценки необходимо иметь в наличии значения функции $f(x)$ в (n+1) узлах, расположенных левее x_n.

Из формул (2.25) и (2.65) следует, что погрешность интерполяции вторым полиномом Ньютона минимальна, если значение аргумента x лежит в интервале (x_{n-1}, x_n).

Запишем без выводов оценки погрешности для интерполяционной формулы Стирлинга

$$| R_S | \approx \frac{\Delta^{2n+1} y_{-n-1} + \Delta^{2n+1} y_{-n}}{2 \cdot (2n+1)!} q(q^2 - 1)\,(q^2 - 4)\,\ldots\,(q^2 - n^2) \qquad (2.66)$$

и интерполяционной формулы Бесселя

$$| R_B | \approx \frac{\Delta^{2n+2} y_{-n-1} + \Delta^{2n+2} y_{-n}}{2 \cdot (2n+2)!} q(q^2 - 1)(q^2 - 4)\ldots(q^2 - n^2)(q - n - 1). \qquad (2.67)$$

В двух последних формулах значения производных тоже выражены через конечные разности.

Видно, что величина погрешности у всех интерполяционных формул уменьшается с ростом количества используемых узлов. Однако в вычислительной математике доказано, что интерполянты, построенные по значениям функции $f(x)$ в равноотстоящих узлах, вообще говоря, обладают плохой сходимостью. Это значит, что в случае произвольной интерполируемой функции $f(x)$ с ростом числа узлов n остаточный член R не обязательно стремится к нулю.

Для оценки погрешности кубического сплайна можно использовать общее выражение (2.57). При задании исходной функции таблицей с постоянным шагом h погрешность ограничена величиной порядка $O(h^4)$ [5, 13].

Следующий принципиальный недостаток, которым обладает интерполяция алгебраическими полиномами, вначале проиллюстрируем следующим примером.

Пример 5. Пусть наши данные о функции $y = f(x)$ состоят из шести пар значений x_i , y_i (i = 0, 1, ..., 5), сведенных в таблицу 2.7.

Таблица 2.7

i	0	1	2	3	4	5
x_i	4,0	4,5	5,5	6,0	7,0	7,5
y_i	3,73	2,37	2,14	0,55	0,09	0,03

Значения функции в таблице 2.7 монотонно убывают. Однако интерполирующий полином Лагранжа (2.8), построенный по данным таблицы 2.7, имеет на интервале [4; 7,5] два явно выраженных максимума и два минимума (см. рис. 2.5).

Завышенные значения интерполянта вблизи x=5 и x=7,3 не обусловлены исходными данными, как и заниженные значения при x=4,4 и x=6,5.

Если же построить интерполянты Лагранжа по первым четырем и последним четырем точкам таблицы 2.7, то мы получим гладкие функции, не обладающие резкими осцилляциями (см. рис. 2.6). Значения новых интерполянтов в точках x = 4,4; 5; 6,5; 7,3 не отличаются сильно от табличных значений y_i в ближайших узлах.

На практике в случаях, когда информация об интерполируемой функции $f(x)$ ограничена только табличными данными вида 2.1, не

используют интерполирующие алгебраические полиномы степени выше 10, даже если точки данных значений x_i; $y_i = f(x_i)$ ложатся на гладкую кривую без резких перегибов. Наиболее широкое распространение имеет кусочно-многочленная интерполяция, пример которой подробно описан в § 2.5. В любых ситуациях интерполянты целесообразно строить по 4 – 6 точкам, окружающим значение аргумента x, для которого требуется вычислить значение функции $y = f(x)$.

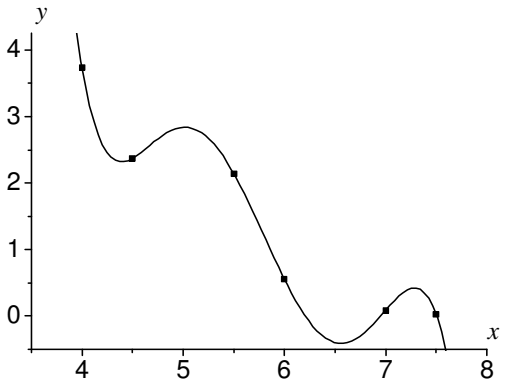
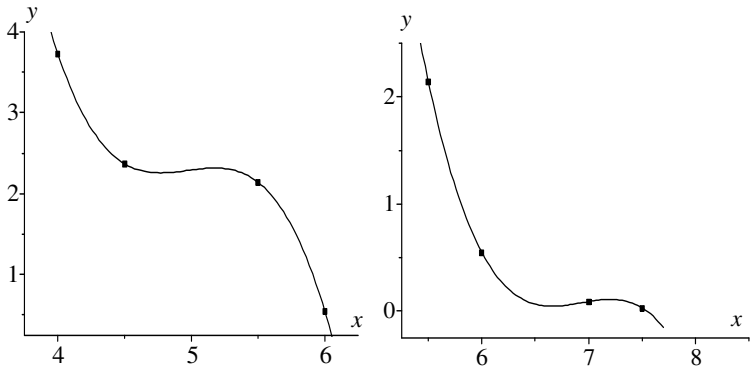


Рис. 2.5. Интерполирующий полином Лагранжа, построенный по данным таблицы 2.7.

Наконец, для практики существенным недостатком любых интерполянтов является их чувствительность к дополнению таблицы исходных данных значениями функции в дополнительных узлах. Пусть по данным таблицы 2.1 (т.е. по n+1 точкам) построен интерполянт n-й степени $P_n(x)$, например, в форме Лагранжа. Если теперь к исходным данным добавить еще одно значение функции $y_{n+1} = f(x_{n+1})$ для дополнительного узла x_{n+1} и вновь построить интерполирующий полином $P_{n+1}(x)$ (уже по n+2 точкам), то он будет существенно отличаться от предыдущего интерполянта. Вычисления значений $P_n(x)$ и $P_{n+1}(x)$ для одного и того же аргумента x (не совпадающего ни с одним узлом) могут давать значения, различающиеся на порядок.

Вышеперечисленные недостатки интерполяции приводят к тому, что в физических задачах для приближения функций чаще используют методы аппроксимации, которые изложены в следующей главе.



а б

Рис.2.6. Интерполирующие полиномы Лагранжа, построенные по фрагментам таблицы 2.7:
а) $4 \le x \le 6$, б) $5{,}5 \le x \le 7{,}5$.

Глава 3. АППРОКСИМАЦИЯ ДАННЫХ

§ 3.1. Проблема аппроксимации

Пусть нашей задачей является исследование зависимости двух непрерывных переменных величин x и y. Явный вид функциональной зависимости $y = f(x)$ неизвестен. В нашем распоряжении имеется конечное число значений исследуемой функции y_i в определенных узлах x_i (i=1,…, n), причем все узлы лежат внутри ограниченного интервала [a, b]. Нам требуется метод расчета значения функции $y = f(x)$ для произвольного аргумента x из интервала [a, b].

В такой постановке задача совпадает с проблемой, поставленной в начале предыдущей главы, которая решается методами интерполяции. В качестве дополнительной информации сформулируем следующие условия.

Во-первых, в нашем распоряжении, кроме таблицы вида 2.1, имеются некоторые теоретические соображения. Предварительные исследования позволили выяснить класс, которому принадлежит функция $f(x)$. Например, физическая теория установила, что зависимость $y = f(x)$ должна быть линейной, т.е. представляться формулой $y = c_1 x + c_0$. Функции $f(x)$, описывающие взаимосвязь физических величин x и y, могут быть полиномиальными, гармоническими, экспоненциальными и т.д.

Во-вторых, данные значения функции y_i (i = 1,…, n) не являются абсолютно точными, а содержат некоторые погрешности ξ_i (i = 1,…, n)

$$y_i = f(x_i) + \xi_i \quad (i = 1, \dots, n). \tag{3.1}$$

Если значения y_i (i = 1, … , n) были получены в эксперименте, то они неизбежно содержат приборную и случайную погрешности измерения.

Приведем несколько примеров из физики.

1. Металлический проводник. Для подобных проводников выполняется закон Ома

$$I = U / R.$$

Иначе говоря, сила тока I прямо пропорциональна приложенной разности потенциалов U. Значение сопротивления R неизвестно. В ходе исследования проведено n измерений силы тока I_i при различных

напряжениях U_i (i = 1, … , n). Измеренные значения содержат случайные погрешности.

2. Затухающие колебания математического маятника. Из теории известно, что амплитуда A уменьшается со временем t по экспоненциальному закону

$$A = A_0 \exp(-k\,t).$$

Коэффициент затухания k неизвестен. Для его вычисления были проведены n измерений амплитуды A_i в моменты времени t_i (i = 1, … , n).

В вышеприведенных примерах задача свелась к необходимости вычисления числового параметра.

Опыт показывает, что имея n точек экспериментальных данных x_i, y_i (i = 1, … , n) и зная класс, которому принадлежит функция $f(x)$, как правило, невозможно подобрать параметр функции так, чтобы график функции $f(x)$ точно прошел через все экспериментальные точки. Например, из практики известно, что при исследовании омического проводника n пар измеренных значений тока и напряжения не ложатся строго на прямую линию с каким-либо углом наклона. Это объясняется тем, что измеренные значения содержат погрешности, как было уже указано в (3.1). Следовательно, приближение исследуемой функции при наличии погрешностей исходных данных не требует строгого выполнения условий (2.1).

Задача аппроксимации ставится следующим образом. Даны n пар значений аргумента и функции: x_i, y_i (i = 1, … , n), полученных, как правило, в эксперименте. Кроме того, из теории известен общий вид функции

$$y = f(x, A, B, C, \ldots), \qquad (3.2)$$

которая связывает исследуемые переменные величины x и y. Функция (3.2) содержит конечное число неизвестных постоянных параметров A, B, C, … . Требуется определить эти численные значения, тогда функция (3.2) будет искомым приближением исследуемой зависимости переменных величин x и y. Зная явный вид такой функции, можно вычислять ее значения для любых аргументов, в т.ч. для $x \notin x_i$ (i = 1, … , n).

Для вычисления параметров функции A, B, C, … должны использоваться экспериментальные данные x_i, y_i (i = 1, … , n). Так как результаты измерений обязательно содержат погрешности, то

рассчитанные значения параметров функции *A, B, C,* … будут обязательно приближенными.

Такая функция вида (3.2), содержащая приближенные значения параметров *A, B, C,* … , полученных с помощью экспериментальных данных x_i, y_i (i = 1, … , n), называется ***аппроксимирующей*** или ***аппроксимантом***.

Методы получения аппроксимирующих функций называются ***аппроксимацией.***

Задачи аппроксимации возникают, например, в ситуациях, когда теория устанавливает вид функциональной зависимости между исследуемыми величинами x и y, но не в состоянии априори получить значения параметров этой функции. Тогда проводятся экспериментальные исследования зависимости величин x и y, и с помощью n пар измеренных значений x_i, y_i (i = 1, … , n) вычисляются искомые параметры аппроксимирующей функции. Для второго из вышеприведенных примеров теория дает экспоненциальное затухание амплитуды с течением времени, но теоретический расчет коэффициента затухания для конкретной колебательной системы может быть весьма сложным и даже практически невозможным. Гораздо проще провести эксперимент и измерить значения уменьшающейся амплитуды. Ниже будет приведен простой метод вычисления коэффициента затухания по данным измерений.

Полученные приближенные значения параметров означают, что методы аппроксимации дают нам не истинную функцию $y = f(x)$, связывающую исследуемые физические величины x и y, а некоторое ее приближение. Следовательно, проблема аппроксимации сводится к методике нахождения значений параметров *A, B, C,* …, которые обеспечивают «наилучшее» приближение аппроксимирующей функции к истинной зависимости. Критерий «наилучшего» приближения базируется на минимизации отклонений значений построенной функции $f(x_i, A, B, C, \ldots)$ в узлах x_i от соответствующих чисел y_i.

В качестве возможной оценки качества аппроксимации можно взять максимальное значение из модулей разностей

$$\max_{i=1,\ldots,n} |f(x_i, A, B, C, \ldots) - y_i|.$$

Можно в качестве критерия «наилучшего» приближения выбрать среднее арифметическое абсолютных значений отклонений $|f(x_i, A, B,$

$C,\ldots) - y_i|$ или среднеквадратичное отклонение. Критерий, наиболее широко используемый для построения аппроксимантов, излагается ниже в данной главе.

Подчеркнем еще раз различие между задачами интерполяции и аппроксимации. Интерполянт должен удовлетворять условию (2.1), а его график – проходить через все точки исходных данных, сведенных в таблицу 2.1. Для аппроксиманта требование (2.1) не обязательно, но должен выполнятся критерий наилучшего приближения, описанный в следующем параграфе.

§ 3.2. Метод наименьших квадратов

На практике, как правило, уменьшение отклонения $|y_i - f(x_i, A, B, C,\ldots)|$ для определенного i-го узла сопровождается увеличением для другого. Практически никогда не удается уменьшить до нуля все абсолютные значения разностей $|f(x_i, A, B, C,\ldots) - y_i|$. Поэтому при построении аппроксиманта вначале необходимо однозначно сформулировать критерий наилучшего приближения аппроксимирующей функции к экспериментальным данным x_i , y_i (i = 1,..., n).

В теории численных методов используются различные критерии аппроксимации. Например, хорошо разработана методика так называемого равномерного приближения ортогональными функциями. Но наибольшее распространение в практике получил подход, называемый ***методом наименьших квадратов***.

После выбора класса аппроксимирующей функции $f(x_i, A, B, C,\ldots)$, которая содержит один или несколько параметров $A, B, C, \ldots$, строится сумма следующего вида:

$$Q = \sum_{i=1}^{n} \left[f(x_i, A, B, C,...) - y_i \right]^2 , \qquad (3.3)$$

где суммирование проводится по всем индексам i = 1, ... n, т.е. по номерам пар исходных данных (измеренных значений) x_i, y_i.

Искомыми значениями параметров $A, B, C, \ldots$ полагаются числа, которые обеспечивают минимум суммы Q. Численное значение величины Q существенно зависит от параметров $A, B, C, \ldots$. Изменяя

величины параметров *A, B, C,* … можно добиться уменьшения суммы до ее минимально возможного значения. Следовательно, на этапе поиска минимизирующих значений параметров *A, B, C,* … сумма *Q* рассматривается как функция переменных *A, B, C,* … . Исходные данные x_i, y_i (i = 1,..., n) являются, естественно, постоянными числами.

Согласно известной теореме математического анализа, необходимым условием экстремума функции нескольких аргументов является равенство нулю первых частных производных по всем аргументам. Следовательно, необходимо приравнять нулю выражения частных производных величины *Q*:

$$\partial Q \big/ \partial A = 0, \quad \partial Q \big/ \partial B = 0, \quad \partial Q \big/ \partial C = 0, \ldots . \tag{3.4}$$

При этом мы получаем систему уравнений для нахождения искомых значений параметров *A, B, C*, … . Важно, что количество уравнений равно количеству неизвестных. В некоторых случаях, например если аппроксимирующей функцией является полином с действительными коэффициентами, решение системы (3.4) не представляет принципиальных трудностей (см. следующий параграф). Напротив, иногда вид аппроксимирующей функции таков, что уравнения (3.4) оказываются нелинейными и весьма трудными для решения. Таким образом, задача аппроксимации приводит к проблеме решения систем как линейных, так и нелинейных уравнений, чему и посвящены соответствующие главы данного учебного пособия.

Заметим, что уравнения (3.4), строго говоря, являются условиями экстремума, но не обязательно минимума величины *Q*. Необходимые условия минимума сформулированы в математическом анализе и требуют исследования вторых частных производных функции *Q* (*A, B, C*, …) по переменным *A, B, C*, … . Однако из способа построения суммы (3.3) непосредственно следует, что любое удаление значений аппроксимирующей функции f (x, *A, B, C*,…) от данных y_i (i = 1, . . . n) приводит к неограниченному возрастанию величины *Q*. Поэтому найденный решением уравнений (3.4) экстремум будет искомым минимумом.

§ 3.3. Аппроксимация алгебраическими полиномами

Задача вычисления минимизирующих значений параметров *A, B, C,* … становится особенно простой, если аппроксимирующая функция выбирается из множества алгебраических полиномов с действительными коэффициентами. В этом случае в выражении (3.3) для суммы Q функция общего вида $f\ (x_i\ ,\ A,\ B,\ C,\ldots)$ заменяется на полином степени m:

$$P_m(x) = \sum_{j=0}^{m} a_j \cdot x^j\ , \tag{3.5}$$

где $m \le n - 1$. Напомним, что n – число узлов таблицы исходных данных x_i, y_i $(i = 1, \ldots n)$.

Теперь задача сводится к вычислению таких коэффициентов полинома a_j, которые минимизируют сумму (3.3) для аппроксимирующей функции вида (3.5):

$$Q = \sum_{i=1}^{n} \left[P_m(x_i) - y_i\right]^2\ . \tag{3.6}$$

Суммирование проводится по индексам $i = 1, \ldots n$, т.е. по номерам пар исходных данных x_i, y_i.

Полином m-й степени содержит $m + 1$ коэффициентов. Для их нахождения сначала полагают величины a_j $(j = 0, \ldots , m)$ переменными. Чтобы найти минимизирующие значения коэффициентов искомого полинома, необходимо взять первые частные производные от величины Q по всем переменным a_j $(j = 0, \ldots , m)$ и приравнять их нулю. Получится система из $m + 1$ уравнений

$$\frac{\partial Q}{\partial a_k} = 0, \tag{3.7}$$

где $k = 0, 1, \ldots, m$ – номер уравнения.

Полученная система уравнений (3.7) является линейной относительно неизвестных a_j − коэффициентов, минимизирующих значение полинома Q. Распишем эту систему в явной форме:

$$\sum_{i=1}^{n}\left[y_i - \sum_{j=0}^{m} a_j x_i^j \right] x_i^k = 0, \qquad k = 0, 1, \ldots, m. \quad (3.8)$$

В системе (3.8) индекс k нумерует уравнение, а индекс j – номер коэффициента a_j при аргументе степени j в полиноме (3.5).

Во всех уравнениях системы (3.8) сменим порядок суммирования и преобразуем эту линейную систему к стандартному виду:

$$\sum_{j=0}^{m}\left[a_j \sum_{i=1}^{n} x_i^{j+k} \right] = \sum_{i=1}^{n}\left[y_i x_i^k \right]. \quad (3.9)$$

Вычисление значений коэффициентов a_j реализуется одним из обычных методов решения систем линейных уравнений, которые рассматриваются в следующей главе.

Аппроксимация данных x_i, y_i (i = 1, . . . n) линейной функцией подробно описана во множестве учебников, учебных и методических пособий. Для расчета аппроксимирующего полинома 1-й степени в линейной системе (3.9) следует положить m = 1. Решение полученной системы двух уравнений элементарно, поэтому сразу запишем результат:

$$P_1(x) = a_1 x + a_0, \quad (3.10)$$

где
$$a_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2},$$

$$a_0 = \frac{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}. \quad (3.11)$$

Теперь рассмотрим не менее важный для практики случай, когда в качестве аппроксимирующего полинома используется квадратичная функция. Для этого, прежде всего, положим в полиноме (3.5) степень m=2:

$$P_2(x) = a_2 x^2 + a_1 x + a_0. \quad (3.12)$$

Аппроксимирующий полином содержит три неопределенных коэффициента. Сумма (3.6) примет следующий вид:

$$Q = \sum_{i=1}^{n} \left[a_2 x_i^2 + a_1 x_i + a_0 - y_i\right]^2.$$

Система (3.7) в этом случае состоит из трех уравнений:

$$\frac{\partial Q}{\partial a_0} = 0, \quad \frac{\partial Q}{\partial a_1} = 0, \quad \frac{\partial Q}{\partial a_2} = 0.$$

Вычислим производные и преобразуем систему к стандартному виду:

$$\begin{aligned}
&a_0 n + a_1 \sum_{i=1}^{n} x_i + a_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i, \\
&a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 + a_2 \sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} x_i y_i, \\
&a_0 \sum_{i=1}^{n} x_i^2 + a_1 \sum_{i=1}^{n} x_i^3 + a_2 \sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} x_i^2 y_i.
\end{aligned} \quad (3.13)$$

Ввиду малого размера последней системы для ее решения можно использовать метод Крамера (см. главу 4).

Для компактной записи решения полученной системы введем обозначения для семи сумм:

$$S_X = \sum_{i=1}^{n} x_i\,, \quad S_{2X} = \sum_{i=1}^{n} x_i^2\,, \qquad S_{3X} = \sum_{i=1}^{n} x_i^3\,, \qquad S_{4X} = \sum_{i=1}^{n} x_i^4\,,$$

$$S_Y = \sum_{i=1}^{n} y_i\,, \quad S_{XY} = \sum_{i=1}^{n} x_i y_i\,, \quad S_{2XY} = \sum_{i=1}^{n} x_i^2 y_i\,.$$

Составим 4 детерминанта:

$$\Delta_Z = \begin{vmatrix} n & S_X & S_{2X} \\ S_X & S_{2X} & S_{3X} \\ S_{2X} & S_{3X} & S_{4X} \end{vmatrix}, \quad \Delta_0 = \begin{vmatrix} S_Y & S_X & S_{2X} \\ S_{XY} & S_{2X} & S_{3X} \\ S_{2XY} & S_{3X} & S_{4X} \end{vmatrix},$$

$$\Delta_1 = \begin{vmatrix} n & S_Y & S_{2X} \\ S_X & S_{XY} & S_{3X} \\ S_{2X} & S_{2XY} & S_{4X} \end{vmatrix}, \quad \Delta_2 = \begin{vmatrix} n & S_X & S_Y \\ S_X & S_{2X} & S_{XY} \\ S_{2X} & S_{3X} & S_{2XY} \end{vmatrix}.$$

Согласно методу Крамера, искомые коэффициенты представляются следующими отношениями детерминантов:

$$a_0 = \Delta_0 / \Delta_Z\,, \quad a_1 = \Delta_1 / \Delta_Z\,, \qquad a_2 = \Delta_2 / \Delta_Z\,. \tag{3.14}$$

Пример 1. Некоторая функция задана таблицей 3.1, содержащей 10 пар значений x_i ; y_i (i = 1, ..., 10).

Таблица 3.1

x_i	1	2	3	4	5	6	7	8	9	10
y_i	0,9	4,2	5,9	7,2	7,6	7,4	6,9	6,2	4,5	2,4

Очевидно, что данная функция $y = f(x)$ является немонотонной. Построим для нее аппроксимирующий полином 2-й степени вида (3.12).

Используя данные таблицы 3.1, вычислим коэффициенты (3.14) этого аппроксимирующего полинома $P_2(x)$:

$$a_0 = -1{,}6467\ ;\quad a_1 = 3{,}3136\ ;\quad a_2 = -0{,}2924.$$

График полученного полинома $P_2(x)$ приведен на рис. 3.1. В тех же осях изображены данные таблицы 3.1 и, для сравнения, график аппроксимирующего полинома 1-й степени P_1 (x), определенный линейной функцией (3.10). Его коэффициенты, рассчитанные по формулам (3.11), равны $a_0 = 4{,}787$; $a_1 = 0{,}097$.

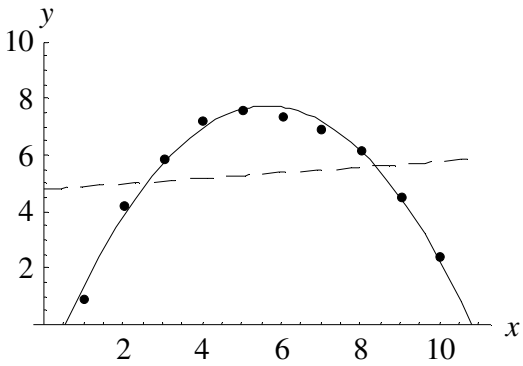


Рис. 3.1. Аппроксимация данных таблицы 3.1 полиномами. Точки – исходные данные, сплошная линия – аппроксимирующий полином 2-й степени, штриховая – аппроксимирующий полином 1-й степени.

Простейшей оценкой качества аппроксимации является величина Q, рассчитанная по выражению (3.6). Чем ближе располагается график аппроксиманта к точкам исходных данных, тем меньше величина Q. Для данных таблицы 3.1 и полученного выше полинома $P_2(x) = -0{,}2924x^2 + 3{,}3136x - 1{,}6467$ величина $Q = 0{,}73$. Для тех же данных и линейного аппроксиманта $P_1(x) = 0{,}097\ x + 4{,}787$ численное значение $Q = 45{,}88$. Столь сильное различие величин Q убедительно иллюстрируется графиками на рис. 3.1. Приведенный пример наглядно убеждает, что для приближения немонотонных функций линейная

аппроксимация дает очень большую погрешность. Требуется аппроксимация нелинейными функциями.

При использовании аппроксимирующих полиномов степени выше второй решение линейной системы уравнений (3.9) целесообразно реализовать методом Гаусса, который описан в следующей главе.

§ 3.4. Аппроксимация суммами Фурье

Частным, но важным случаем является аппроксимация табличных данных суммами тригонометрических функций с кратными периодами. Такие функции в математике получили название сумм Фурье. Аппроксимация суммами Фурье применяется, в частности, в гармоническом анализе при исследовании различных периодических функций. В физике эти суммы используются при исследовании периодических процессов и кристаллических структур.

Пусть, как и в предыдущих задачах аппроксимации, функция задана набором n пар значений

$$x_i, y_i = f(x_i) \quad (i = 1, ..., n), \tag{3.15}$$

причем все узлы x_i ($i = 1, ... n$) лежат внутри конечного интервала $[a, b]$ длиной L. С помощью линейного преобразования аргумента точки узлов всегда можно привести к интервалу $(0, 2\pi)$. Определим новую переменную

$$t = 2\pi(x - a)/L \tag{3.16}$$

и числа x_i ($i = 1, ..., n$) по формуле (3.16) пересчитаем в соответствующие значения t_i. Все новые значения узлов t_i ($i = 1, ..., n$) будут принадлежать интервалу $(0, 2\pi)$.

Аппроксимирующую функцию выберем в виде суммы Фурье:

$$T_m(t) = a_0 + \sum_{k=1}^{m} [a_k \cos(kt) + b_k \sin(kt)]. \tag{3.17}$$

Количество слагаемых m в сумме (3.17) будем называть степенью или порядком суммы Фурье. Число m можно выбирать произвольным (конечно, целым) при выполнении условия $n \geq 2m + 1$.

Найдем значения коэффициентов a_0, a_k, b_k, ($k = 1, ..., m$) для суммы Фурье (3.17), дающие наилучшее приближение к табличным данным (3.15) согласно критерию наименьших квадратов. Решение поставленной задачи реализуется путем, аналогичным изложенному в предыдущем параграфе. Сначала в функции $f(x)$ проведем замену аргумента согласно (3.16), а затем в выражении (3.3) преобразованную функцию $f(t)$ заменяем на сумму Фурье (3.17).

Будем подбирать такие значения коэффициентов a_0, a_k, b_k ($k = 1, ... , m$), которые обеспечат минимум суммы Q (3.3). Для этого возьмем от величины Q частные производные по переменным a_0, a_k, b_k ($k = 1, ..., m$) и приравниваем их к нулю. При этом получается система из $2m + 1$ уравнений для $2m + 1$ неизвестных коэффициентов суммы (3.17). Аппроксимирующая функция имеет такой вид, что полученная система уравнений оказывается линейной относительно $2m + 1$ искомых величин a_0, a_k, b_k ($k = 1, ..., m$):

$$n\,a_0 + \sum_{k=1}^{m} \{a_k \sum_{i=1}^{n} \cos(kt_i) + b_k \sum_{i=1}^{n} \sin(kt_i)\} = \sum_{i=1}^{n} y_i\,, \qquad (3.18)$$

$$a_0 \sum_{i=1}^{n} \cos(jt_i) + \sum_{k=1}^{m} \{a_k \sum_{i=1}^{n} \cos(kt_i)\cos(jt_i) + b_k \sum_{i=1}^{n} \sin(kt_i)\cos(jt_i)\} = \sum_{i=1}^{n} y_i \cos(jt_i)\,,\ j = 1, ..., n, \qquad (3.19)$$

$$a_0 \sum_{i=1}^{n} \sin(jt_i) + \sum_{k=1}^{m} \{a_k \sum_{i=1}^{n} \cos(kt_i)\sin(jt_i) + b_k \sum_{i=1}^{n} \sin(kt_i)\sin(jt_i)\} = \sum_{i=1}^{n} y_i \sin(jt_i)\,,\ j = 1, ... , n. \qquad (3.20)$$

Методы решения систем линейных уравнений изложены в следующей главе, поэтому полагаем, что в принципе поставленная задача решена.

Величина a_0 является постоянной составляющей аппроксимирующей функции. Параметры a_k и b_k (k = 1, ..., m) представляют собой амплитуды соответственно четных и нечетных гармоник Фурье-суммы.

Вычисленные значения параметров a_0, a_k, b_k (k = 1, ..., m) полностью определяют искомую аппроксимирующую функцию (3.17). Для того чтобы вернуться к первоначальному аргументу x, достаточно использовать линейное преобразование, обратное преобразованию (3.16):

$$x = t\,L/(2\pi) + a. \tag{3.21}$$

При этом аппроксимирующая Фурье-сумма может быть записана в следующем виде:

$$T_m(x) = a_0 + \sum_{k=1}^{m} [a_k \cos(2\pi k(x-a)/L) + b_k \sin(2\pi k(x-a)/L)]\,. \tag{3.22}$$

Полученный аппроксимант $T(x)$ является периодической функцией с периодом L, иначе говоря, для любого аргумента x в его области определения выполняется равенство $T(x+L) = T(x)$.

Иногда при исследовании функции f (x) имеется возможность выбрать узлы x_i (i = 1, ..., n) равноотстоящими. Тогда преобразование (3.16) даст узлы t_i (i = 1, ..., n), равноотстоящие на интервале (0, 2π):

$$t_i = i\frac{2\pi}{n} \qquad (i = 1, ..., n). \tag{3.23}$$

В этих случаях многие суммы в уравнениях (3.18) – (3.20) зануляются и эта система упрощается. Выражения для искомых коэффициентов можно записать в следующем явном виде:

$$a_0 = \frac{1}{n}\sum_{i=1}^{n} y_i\,, \tag{3.24}$$

$$a_k = \frac{2}{n}\sum_{i=1}^{n} y_i \cos(kt_i)\,, \quad k = 1, ..., m, \tag{3.25}$$

$$b_k = \frac{2}{n} \sum_{i=1}^{n} y_i \sin(kt_i) \ , \quad k = 1, \ldots m. \tag{3.26}$$

Вычисленные значения коэффициентов a_0, a_k, b_k (k = 1, ..., m) подставляются в общее выражение Фурье-аппроксиманта (3.22).

Пример 2. Функция f (x) задана таблицей 3.2, содержащей 10 пар значений x_i, y_i (i = 1, ..., 10).

Таблица 3.2

x_i	1	2	3	4	5	6	7	8	9	10
y_i	1,3	2,8	2,5	1,1	–0,8	–1,9	–1,1	0,9	3,85	0

Проведем аппроксимацию функции f (x) Фурье-суммой вида (3.22) на интервале [0, 10]. Сначала вычислим значения узлов t_i (i = 1, ..., 10) по формуле (3.16), используя a = 0 и L = 10. Так как в данной задаче узлы равноотстоящие, то коэффициенты Фурье-аппроксиманта рассчитываются по формулам (3.24) – (3.26). Ограничимся количеством слагаемых Фурье-суммы m = 6. В результате получается a_0 = 0,86, остальные коэффициенты сведены в таблицу 3.3.

Таблица 3.3

k	1	2	3	4	5	6
a_k	1,257	–0,719	–0,577	–0,541	–0,56	–0,541
b_k	1,105	–1,246	–0,551	–0,323	0	0,551

Вычисленные коэффициенты подставляются формулу (3.22) с учетом значений параметров a = 0 и L = 10:

$$T_m(x) = a_0 + \sum_{k=1}^{m} [a_k \cos(\pi kx/5) + b_k \sin(\pi kx/5)] . \tag{3.27}$$

На рис. 3.2 приведены графики Фурье-сумм (3.27) для степеней m=2 и m=6.

На выбор степени Фурье-аппроксиманта влияют в значительной мере погрешности аппроксимируемых данных x_i, y_i (i = 1, ..., n).

Например, может быть, что в примере 2 данные y_i (i = 1, ..., n) получены с большой погрешностью. Тогда значение y_9 в таблице 3.2 может быть сильно завышено по сравнению с истинной величиной, а значение y_{10}, наоборот, занижено. В этом случае исходные данные удовлетворительно аппроксимируются суммой Фурье низкого порядка m =2 (см. рис. 3.2).

В противоположном случае, если погрешность полученных значений y_i (i = 1, ..., n) мала, то график аппроксиманта должен проходить вблизи точек исходных данных. Для этого степень Фурье-аппроксиманта приходится повышать, что приводит, в свою очередь, к появлению дополнительных перегибов исследуемой функции f (x), достоверность наличия которых приходится обосновывать. В частности, перегибы аппроксиманта степени m =6 в областях около значений x = 1,5; x = 3,5; x = 5,5; x=7,5 на рис.3.2 явно не следуют из данных таблицы 3.2. Таким образом, задача выбора степени Фурье-аппроксиманта не может быть однозначно решена, если ограничиваться формальными рамками численных методов.

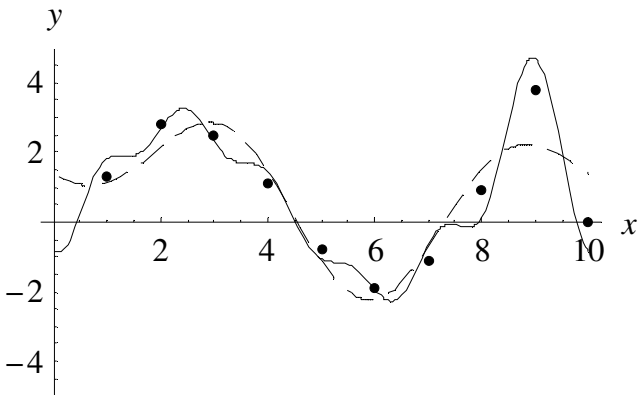


Рис. 3.2. Аппроксимация данных суммой Фурье.
Точки – исходные данные, сплошная линия – Фурье-аппроксимант с m=6, штриховая – Фурье-аппроксимант с m=2.

Окончательный выбор порядка Фурье-аппроксиманта делается на основе дополнительной информации об исходных данных, прежде всего

с помощью теоретических соображений о виде исследуемой функции $f(x)$.

§ 3.5. О нелинейной аппроксимации

К нелинейной аппроксимации прибегают в специальных случаях, когда вид нелинейной зависимости $y = f\ (x,\ A,\ B,\ C,\ldots)$ между двумя физическими величинами x и y надежно установлен, но численные значения постоянных параметров $A,\ B,\ C,\ldots$ неизвестны.

Некоторые часто встречающиеся в физике функции сводятся к рассмотренным выше полиномам путем определенного алгебраического преобразования. В частности, зависимость из примера 2 §3.1 логарифмированием превращается в линейную:

$$ln(A) = ln(A_0) - k\,t. \tag{3.28}$$

Пусть в результате эксперимента были получены n значений амплитуд A_i для значений времени t_i $(i = 1, \ldots, n)$. Введем обозначения

$$z = ln(A) \text{ и } a = ln(A_0)$$

и по данным измерений вычислим значения $z_i = ln(A_i)$ для всех $i = 1, \ldots, n$. Тогда следует рассматривать числа t_i, z_i $(i = 1, \ldots, n)$ как значения аргумента и функции соответственно. По формулам (3.11) легко подсчитать параметры линейного аппроксиманта (3.10), при этом свободным членом будет величина a, угловым коэффициентом – искомый коэффициент затухания k с обратным знаком. Если потребуется получить начальную амплитуду A_0, то для этого достаточно вычислить $A_0 = \exp(a)$.

К сожалению, во многих других случаях метод наименьших квадратов приводит к необходимости решать систему нелинейных уравнений.

Пример 3. Весьма распространенной функцией, описывающей различные физические процессы, является следующая:

$$y = A_0\,(1 - e^{-b\,x}), \tag{3.29}$$

где A_0 и b – постоянные параметры.

Пусть числа A_0 и b нам неизвестны, но в нашем распоряжении имеется таблица значений x_i, y_i ($i = 1, \ldots, n$) вида (3.1). Действуя методом наименьших квадратов, составляем величину Q, согласно (3.3), затем дифференцируем функцию Q (A_0 , b) по переменным A_0 и b, приравниваем частные производные нулю $\partial Q / \partial A_0 = 0$, $\partial Q / \partial b = 0$ и получаем уравнения:

$$A_0 \sum_{i=1}^{n} (1-\exp(-bx_i))^2 - \sum_{i=1}^{n} y_i (1-\exp(-bx_i)) = 0,$$

$$A_0 \sum_{i=1}^{n} (1-\exp(-bx_i)) x_i \exp(-bx_i) - \sum_{i=1}^{n} x_i y_i \exp(-bx_i) = 0. \qquad (3.30)$$

Выражая величину A_0 из первого уравнения и подставляя ее во второе, получим довольно громоздкое нелинейное уравнение для нахождения неизвестного b

$$\sum_{i=1}^{n} y_i (1-\exp(-bx_i)) \times \sum_{i=1}^{n} (1-\exp(-bx_i)) x_i \exp(-bx_i) -$$

$$- \sum_{i=1}^{n} x_i y_i \exp(-bx_i) \times \sum_{i=1}^{n} (1-\exp(-bx_i))^2 = 0. \qquad (3.31)$$

После нахождения величины b коэффициент A_0 легко вычисляется из любого уравнения системы (3.30).

Таким образом, нелинейная аппроксимация, как правило, требует применения методов решения нелинейных уравнений, которые рассматриваются в главе 7 настоящего учебного пособия.

Пример 4. Для решения многих прикладных задач используется функция Гаусса, которую можно записать в следующем виде:

$$y = A \exp\left[-\left(\frac{x-b}{c} \right)^2 \right]. \qquad (3.32)$$

Функция (3.32) содержит три постоянных параметра: A, b и c. Зависимость вида (3.32) используется, например, для анализа различных спектральных характеристик, получаемых экспериментальным путем. Каждый максимум спектра аппроксимируется функцией вида (3.32). Параметры A, b и c для каждого максимума имеют определенный физический смысл.

Для вычисления этих параметров используется таблица данных значений x_i, y_i ($i = 1, \ldots, n$) и описанный выше метод наименьших квадратов. В общее выражение (3.3) для величины Q подставляется функциональная зависимость (3.32). Читателям предлагается самостоятельно записать явный вид функции Q (A, b, c) для данного примера, составить уравнения $\partial Q/\partial A = 0$, $\partial Q/\partial b = 0$, $\partial Q/\partial c = 0$ и оценить сложность их решения.

В качестве комментария к главам 2 и 3 отметим особенности современной терминологии. Во многих иностранных учебных пособиях и соответствующих программных пакетах под интерполяцией и аппроксимацией понимается практически одно и то же. Разница заключается в том, что при аппроксимации функция выдается в явном виде, а при интерполяции вычисляется значение функции при аргументе, отличном от узла интерполяции. Для численных методов, рассмотренных в главе 3 данной книги, часто используется термин «регрессия».

В программах Origin и Mathematica для вычисления аппроксимирующей функции применяется процедура под названием Fit, что в буквальном переводе с английского означает «подгонка».

Несмотря на распространение новых программных продуктов, авторы предпочли в этой книге сохранить классическую терминологию, используемую в российских курсах численных методов. Авторы полагают, что читатель, знакомый с алгеброй и основами математического анализа, с помощью настоящего учебного пособия самостоятельно разберется в смысле и назначении предлагаемых численных методов.

Глава 4. РЕШЕНИЕ СИСТЕМ ЛИНЕЙНЫХ УРАВНЕНИЙ

§ 4.1. Системы линейных уравнений

Решение множества физических задач приводит к системам линейных уравнений относительно нескольких неизвестных. Часто системы линейных уравнений возникают в ходе решения различных задач прикладной математики. В качестве частных примеров можно привести интерполяцию сплайнами или аппроксимацию полиномами, рассмотренные в предыдущих главах данного пособия. Существуют эффективные методы решения дифференциальных уравнений в частных производных, которые сводятся к решению систем линейных уравнений большой размерности.

Запишем в общем виде систему n линейных уравнений для n неизвестных x_i, где $i = 1, 2, ..., n$:

$$\begin{aligned} &a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n = b_1 \\ &a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n = b_2 \\ &\quad ..... \\ &a_{n1}x_1 + a_{n2}x_2 + ... + a_{nn}x_n = b_n \end{aligned} \qquad (4.1)$$

или в краткой форме: $\sum_{j=1}^{n} a_{ij} \cdot x_j = b_i \qquad (i = 1, 2, ..., n).$

Коэффициенты a_{ij} и b_i ($i, j = 1, 2, ..., n$) являются известными величинами. Первый индекс у коэффициента a_{ij} означает номер уравнения в системе, второй индекс – номер неизвестного, на которое умножается данный коэффициент a_{ij}. Коэффициенты b_i в правых частях уравнений называются ***свободными членами***.

Решить систему (4.1) означает отыскать числа x_i ($i = 1, 2, ..., n$), которые при подстановке в (4.1) обращают все уравнения в тождества. Набор этих чисел называется ***корнями*** системы (4.1).

Коэффициенты при неизвестных образуют квадратную матрицу **A** размером $n \times n$:

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}. \qquad (4.2)$$

Свободные члены и корни представляются столбцами или n-мерными векторами $\boldsymbol{b}$ и $\boldsymbol{x}$. Тогда систему (4.1) можно записать более кратко в матричном представлении:

$$\boldsymbol{A} \cdot \boldsymbol{x} = \boldsymbol{b} \qquad (4.3)$$

В линейной алгебре доказана следующая теорема: если детерминант (определитель) матрицы $\boldsymbol{A}$ не равен нулю, то система уравнений (4.1) имеет единственное решение. Т.е. при $\det \boldsymbol{A} \neq 0$ существует только один набор чисел x_i (i = 1, 2, ..., n), который обращает в тождество систему (4.1). Следовательно, если мы ищем единственное решение системы уравнений (4.1), то прежде чем ее решать, необходимо убедиться, что детерминант матрицы коэффициентов при неизвестных не равен нулю.

Способам вычисления детерминантов посвящена глава 5.

Многочисленные способы решения системы линейных уравнений принято подразделять на две группы: прямые методы и итерационные. Прямые методы позволяют найти решение с помощью определенного конечного количества операций. В итерационных методах решение ищется как предел последовательных приближений, где число итераций зависит от задаваемой погрешности корней. В §§4.1 – 4.4 рассмотрены простейшие прямые методы, итерационным методам посвящен §4.5.

Предположим, что условие $\det \boldsymbol{A} \neq 0$ для системы (4.1) выполнено. Тогда, согласно известной теореме матричной алгебры, для матрицы $\boldsymbol{A}$ существует обратная матрица $\boldsymbol{A}^{-1}$, для которой справедливы равенства:

$$\boldsymbol{A} \cdot \boldsymbol{A}^{-1} = \boldsymbol{A}^{-1} \cdot \boldsymbol{A} = \boldsymbol{E}, \qquad (4.4)$$

где $\boldsymbol{E}$ – единичная матрица размера $n \times n$.

Умножим слева обе части уравнения (4.3) на обратную матрицу $\boldsymbol{A}^{-1}$. Принимая во внимание (4.4), получим:

$$\boldsymbol{x} = \boldsymbol{A}^{-1} \cdot \boldsymbol{b}. \qquad (4.5)$$

Таким образом, умножение обратной матрицы $\boldsymbol{A}^{-1}$ на вектор (столбец) свободных членов $\boldsymbol{b}$ дает единственное решение: вектор корней $\boldsymbol{x}$ исходной системы (4.1). С точки зрения математики проблема решена.

Однако получение обратной матрицы в явном виде сопряжено с практическими трудностями. С ростом числа n требуемый объем вычислений резко возрастает. Заметим, что умножение матрицы на вектор представляет собой простую задачу, легко решаемую программным путем.

Задаче вычисления обратной матрицы посвящена глава 6. Далее в настоящей главе рассматриваются элементарные методы решения систем линейных уравнений, не требующие вычисления обратной матрицы.

§ 4.2. Метод Крамера

Известно, что обратную матрицу можно представить в виде отношения:

$$\boldsymbol{A}^{-1} = \boldsymbol{A}^{*} / \Delta, \tag{4.6}$$

где $\boldsymbol{A}^{*}$ – матрица союзная исходной $\boldsymbol{A}$, т.е. транспонированная матрица алгебраических дополнений, Δ – детерминант матрицы $\boldsymbol{A}$.

Подставим (4.6) в уравнение (4.5) и умножим матрицу $\boldsymbol{A}^{*}$ слева на столбец свободных членов $\boldsymbol{b}$. В результате получим новый n-мерный вектор, компоненты которого Δ_i выражаются следующими суммами:

$$\Delta_i = \sum_{j=1}^{n} A_{ji} \cdot b_j \,, \tag{4.7}$$

где A_{ij} – алгебраические дополнения элементов a_{ij} матрицы (4.2).

Можно доказать, что каждая величина Δ_i (i = 1, 2, ..., n) равна детерминанту матрицы, которая получается из исходной $\boldsymbol{A}$ заменой i-го столбца на столбец свободных членов $\boldsymbol{b}$:

$$\begin{bmatrix} a_{11} & \dots & a_{1,i-1} & b_1 & a_{1,i+1} & \dots & a_{1n} \\ a_{21} & \dots & a_{2,i-1} & b_2 & a_{2,i+1} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & \dots & a_{n,i-1} & b_n & a_{n,i+1} & \dots & a_{nn} \end{bmatrix}. \qquad (4.8)$$

Для доказательства достаточно непосредственно вычислить детерминант матрицы (4.8) известным методом разложения по минорам *i*-го столбца, который описан в главе 5.

Следовательно, деление каждого числа Δ_i на детерминант Δ матрицы (4.2) даст нам значение *i*-го корня исходной системы (4.1):

$$x_i = \Delta_i / \Delta , \quad i = 1, 2, \dots, n. \qquad (4.9)$$

Описанная процедура вычисления корней системы линейных уравнений называется ***методом Крамера***. В этом методе решение системы (4.1) полностью сводится к вычислению детерминантов, способы расчета которых подробно изложены в следующей главе.

Следует заметить, что в тех случаях, когда детерминант Δ близок к нулю, метод Крамера может дать большую ошибку в значениях найденных корней, что непосредственно следует из формулы (4.9). Конечно, погрешность вычисления корня существенно зависит и от величины соответствующего детерминанта Δ_i, стоящего в числителе формулы (4.9). Системы уравнений (4.1), у которых детерминант матрицы коэффициентов при неизвестных близок к нулю, требуют для своего решения специальных методов, которые не включены в настоящее учебное пособие.

Важнейшим недостатком алгоритма Крамера является большое количество необходимых арифметических операций. Из формул (4.9) ясно, что для решения данной системы (4.1) приходится вычислять (n+1) детерминантов, каждый порядка n. Рациональные методы расчета детерминантов рассматриваются в следующей главе. Пока заметим, что для вычисления детерминанта n-го порядка способом разложения по минорам требуется произвести $n!\cdot(n - 1)$ умножений и примерно столько же делений. Следовательно, при решении системы линейных уравнений 40-го порядка методом Крамера придется выполнить $41\cdot 40!\cdot 39$ умножений и сложений. Так как $40! \approx 8\cdot 10^{47}$, то можно оценить, что при использовании самых быстродействующих компьютеров необходимое время решения во много раз превышает

время существования нашей Вселенной. Кроме того, при выполнении процессором большого количества операций накапливается значительная погрешность результата, так как в каждой операции над реальными числами проводится округление.

При решении систем линейных уравнений для $n = 2$ и $n = 3$ метод Крамера легко реализуется и без использования компьютера. Для больших значений n целесообразно применять более эффективные методы.

§ 4.3. Метод Гаусса

Метод Гаусса отличается сравнительно малым объемом вычислений и простотой алгоритма, что позволяет легко его программировать. Этот метод заключается, в сущности, в последовательном исключении неизвестных.

Процедура вычисления корней системы линейных уравнений состоит из двух частей: прямого хода и обратного хода.

Прямой ход состоит из отдельных шагов.

Первый шаг. Выбирается ***ведущий элемент*** − коэффициент при неизвестном x_1. Это может быть любой коэффициент a_{i1} ($i = 1, ..., n$) не равный нулю. Напомним, что первый индекс у коэффициента a_{ij} означает номер уравнения в системе, второй индекс − номер неизвестного, на которое умножается данный коэффициент a_{ij}. Порядок расположения уравнений в системе (4.1) не существенен. Поэтому, переставляя уравнения, всегда можно ведущим сделать элемент a_{11}.

Очевидно, что все коэффициенты a_{i1} ($i = 1, ..., n$) не могут равняться нулю. Равенство нулю всех a_{i1} ($i = 1, ..., n$) означает наличие в матрице $\boldsymbol{A}$ нулевого столбца и равенство нулю детерминанта матрицы. В этом случае единственное решение системы отсутствует.

Уравнение с ведущим элементом делится на коэффициент $a_{11} \neq 0$ и приобретает вид:

$$x_1 + \sum_{j=2}^{n} \frac{a_{1j}}{a_{11}} x_j = \frac{b_1}{a_{11}} . \qquad (4.10)$$

Теперь уравнение (4.10) умножается на коэффициент a_{21} и вычитается из 2-го уравнения исходной системы (4.1). Слагаемое, содержащее x_1, исчезает из 2-го уравнения.

Затем уравнение (4.10) последовательно умножается на коэффициенты a_{i1} (где $i = 3, ..., n$) и вычитается из соответствующего i-го уравнения исходной системы. В каждом уравнении член, содержащий x_1, исчезнет. В результате, кроме уравнения (4.10), мы получим систему из $n - 1$ уравнений, которая содержит $n - 1$ неизвестных x_i ($i = 2, ... n$):

$$\sum_{j=2}^{n} a_{ij}^{(1)} \cdot x_j = b_i^{(1)}, \quad i = 2, ..., n. \tag{4.11}$$

В системе уравнений (4.11) введены обозначения

$$a_{ij}^{(1)} = a_{ij} - a_{i1}\frac{a_{1j}}{a_{11}}, \quad b_i^{(1)} = b_i - a_{i1}\frac{b_1}{a_{11}}, \quad i,j = 2, ..., n. \tag{4.12}$$

Второй шаг прямого хода выполняется аналогично. В системе уравнений (4.11) выбирается ведущий элемент $a_{i2}^{(1)} \neq 0$ ($i = 2, ..., n$) при x_2. При необходимости меняется порядок уравнений в системе (4.11), и ведущим становится $a_{22}^{(1)}$. Сначала первое уравнение системы (4.11) делится на ведущий элемент $a_{22}^{(1)}$:

$$x_2 + \sum_{j=3}^{n} \frac{a_{2j}^{(1)}}{a_{22}^{(1)}} x_j = \frac{b_2^{(1)}}{a_{22}^{(1)}}. \tag{4.13}$$

Затем уравнение (4.13) последовательно умножается на коэффициенты $a_{i2}^{(1)}$ ($i = 3, ..., n$) и вычитается из соответствующего i-го уравнения системы (9). В результате из всех уравнений исчезают члены с x_2. Получим новую систему из $n - 2$ уравнений, которая содержит $n - 2$ неизвестных x_i ($i = 3, ..., n$):

$$\sum_{j=3}^{n} a_{ij}^{(2)} \cdot x_j = b_i^{(2)}, \quad i = 3, ..., n. \tag{4.14}$$

В системе (4.14) введены новые обозначения:

$$a_{ij}^{(2)} = a_{ij}^{(1)} - a_{i2}^{(1)} \frac{a_{2j}^{(1)}}{a_{22}^{(1)}}, \quad b_i^{(2)} = b_i^{(1)} - a_{i2}^{(1)} \frac{b_2^{(1)}}{a_{22}^{(1)}}, \ i,j = 3, ..., n. \quad (4.15)$$

Продолжение процедуры последовательного исключения неизвестных приведет к тому, что после $(n - 1)$-го шага прямого хода получится линейное уравнение, содержащее только одно неизвестное:

$$a_{nn}^{(n-1)} \cdot x_n = b_n^{(n-1)}. \quad (4.16)$$

Таким образом, исходная система уравнений приведена к эквивалентной (т.е. имеющей те же корни), имеющей треугольную матрицу коэффициентов при неизвестных. При этом были преобразованы и свободные члены.

Заметим, что коэффициенты при неизвестных и свободные члены преобразовывались по идентичным рекуррентным формулам. Поэтому целесообразно предварительно построить ***расширенную матрицу*** размером $n \times (n + 1)$, присоединив справа к квадратной матрице коэффициентов при неизвестных $(n + 1)$-й столбец свободных членов. Тогда элементы расширенной матрицы c_{ij} ($i = 1, ..., n, j = 1, \ldots, (n + 1)$) будут преобразовываться по рекуррентному закону:

$$c_{ij}^{(k)} = c_{ij}^{(k-1)} - c_{ik}^{(k-1)} \frac{c_{kj}^{(k-1)}}{c_{kk}^{(k-1)}}, \quad (4.17)$$

где k – порядковый номер шага прямого хода.

После преобразования коэффициентов при неизвестных и свободных членов реализуется ***обратный ход***.

На первом шаге обратного хода из уравнения (4.16) сразу вычисляется n-й корень системы (4.1):

$$x_n = \frac{b_n^{(n-1)}}{a_{nn}^{(n-1)}}. \quad (4.18)$$

В предпоследнем $(n - 2)$-м шаге прямого хода было получено уравнение

$$x_{n-1} + \frac{a_{n-1,n}^{(n-2)}}{a_{n-1,n-1}^{(n-2)}} \cdot x_n = b_{n-1}^{(n-2)}. \qquad (4.19)$$

Второй шаг обратного хода заключается в подстановке (4.18) в (4.19) и вычислении корня x_{n-1}.

Далее последовательно используются уравнения, полученные ранее на каждом шаге прямого хода. На предпоследнем шаге обратного хода вычисляется x_2 с помощью уравнения (4.13). Значения корней x_i (i = 3, ..., n) уже получены в предыдущих шагах. На последнем (n-м) шаге вычисляется x_1 с помощью уравнения (4.10).

Таким образом, исходная система линейных уравнений (4.1) полностью решена.

Примечание. В линейной алгебре доказано, что если $\det \boldsymbol{A} \neq 0$, то на каждом шаге прямого хода найдется хотя бы один ведущий коэффициент, не равный нулю. Следовательно, можно специально не вычислять предварительно детерминант матрицы коэффициентов при неизвестных. Если на каком-либо шаге прямого хода не найдется ненулевой ведущий коэффициент, то это значит, что $\det \boldsymbol{A} = 0$ и система не имеет единственного решения.

Полезно сравнить количество вычислений, необходимых для решения системы линейных уравнений (4.1) методом Крамера и методом Гаусса, т.е. сравнить эффективности двух рассмотренных методов.

Если детерминанты вычислять методом разложения по минорам (см. главу 5), то для расчета одного детерминанта n-го порядка приходится совершать количество операций умножения и деления, равное $(n-1)\times \times(n^2+n+3)\ /\ 3$. Это значит, что для решения системы из n линейных уравнений методом Крамера необходимо выполнить $(n+1)\ (n-1)\ (n^2+n+3)\ /\ 3 + n$ операций умножения и деления. Иначе говоря, число требуемых операций имеет порядок $\sim n^4$.

Можно вычислить, что при решении системы линейных уравнений n-го порядка методом Гаусса в прямом ходе выполняется $n\ (n+1)\ (n+2)\ /\ 3$ операций умножения и деления и столько же операций вычитания. При обратном ходе выполняется $n\ (n-1)\ /\ 2$ операций умножения и деления и столько же операций вычитания. Следовательно, общее количество арифметических действий при решении системы линейных уравнений n-го порядка методом Гаусса имеет порядок $\sim n^3$.

Таким образом, количество арифметических операций при решении системы линейных уравнений в методе Гаусса на порядок меньше, чем в методе Крамера. Это значит, что уже при $n \geq 4$ целесообразно использовать метод Гаусса.

Проиллюстрируем метод Гаусса примером, причем для экономии места ограничимся системой из трех уравнений.

Пример 1. Дана следующая система уравнений:

$$\begin{aligned} 2x_1 + 2x_2 + 4x_3 &= 18, \\ 2x_1 - x_2 + 3x_3 &= 9, \\ 3x_1 - x_2 + 2x_3 &= 7. \end{aligned}$$

Возьмем в качестве первого ведущего коэффициента $a_{11} = 2$. Поделим на него первое уравнение. Затем исключим x_1 из второго и третьего уравнений вычитанием преобразованного первого, домноженного на 2 и 3 соответственно. В результате первого шага прямого хода система принимает вид:

$$\begin{aligned} x_1 + x_2 + 2x_3 &= 9, \\ -3x_2 - x_3 &= -9, \\ -4x_2 - 4x_3 &= -20. \end{aligned}$$

Второй шаг начинается с деления второго уравнения на $a_{22}^{(1)} = -3$. Затем исключается x_2 из третьего уравнения сложением его с преобразованным вторым, умноженным на 4. Второй (и последний в данном примере) шаг прямого хода заканчивается получением треугольной матрицы коэффициентов при неизвестных:

$$\begin{aligned} x_1 + x_2 + 2x_3 &= 9, \\ x_2 + (1/3)\, x_3 &= 3, \\ -(8/3)\, x_3 &= -8. \end{aligned}$$

Обратный ход начинается с вычисления x_3 из последнего уравнения преобразованной системы: $x_3 = 3$. Подставив полученное значение x_3 во второе уравнение, получаем $x_2 = 2$. Наконец, подстановка значений x_3 и x_2 в первое уравнение дает $x_1=1$.

Подставив найденные значения корней в исходную систему, убеждаемся в правильности найденного решения.

§ 4.4. Уточнение корней и число обусловленности

Погрешность, накапливающаяся при вычислениях, приводит к тому, что любой метод (Гаусса, Крамера и т.д.) дает не точное, а приближенное значение корней решаемой системы линейных уравнений.

Сначала рассмотрим случай, когда исходные данные: коэффициенты матрицы $\boldsymbol{A}$ и свободные члены имеют достаточную точность (например, все значащие цифры являются верными, а погрешность не превышает единицы младшего разряда). В этом случае необходимо оценить погрешность, которая накапливается в процессе выполнения алгоритма.

Для этого после решения системы следует полученные значения корней подставить в исходную систему (4.1) и вычислить левые части уравнений. Если они совпадут с соответствующими свободными членами (в пределах допустимых погрешностей), то решение можно полагать удовлетворительным. Если же наблюдается значимое расхождение, то необходимо применить более точный метод решения системы уравнений (4.1) или использовать процедуру уточнения корней.

Обозначим через $\boldsymbol{x}$ вектор точных значений корней, $\boldsymbol{x}^{(1)}$ – вектор приближенных значений корней, полученных в ходе решения. Тогда точное решение можно представить в виде:

$$\boldsymbol{x} = \boldsymbol{x}^{(1)} + \boldsymbol{\delta}, \tag{4.20}$$

где $\boldsymbol{\delta}$ – вектор погрешностей. Каждая компонента вектора $\boldsymbol{\delta}$ представляет собой разность между точным значением i-го корня и его приближенным значением

$$\delta_i = x_i - x_i^{(1)} \quad (i = 1, \ldots, n). \tag{4.21}$$

Подставляя (4.20) в исходную систему (4.3), получим новую систему линейных уравнений:

$$\boldsymbol{A} \cdot \boldsymbol{\delta} = \boldsymbol{\varepsilon}\,, \tag{4.22}$$

где вектор $\boldsymbol{\varepsilon}$ называется ***невязкой*** приближенного решения

$$\boldsymbol{\varepsilon} = \boldsymbol{b} - \boldsymbol{A} \cdot \boldsymbol{x}^{(1)}. \tag{4.23}$$

Процедура уточнения корней заключается в следующих операциях. Полученные приближенные значения корней $\boldsymbol{x}^{(1)}$ подставляются в (4.23) и вычисляются компоненты вектора невязок $\boldsymbol{\varepsilon}$. Далее решается система уравнений (4.22), которая содержит исходную матрицу коэффициентов при неизвестных (4.2), а в качестве свободных членов используются вычисленные невязки (4.22). Решение системы проводится любым доступным способом, например тем же методом Гаусса. Результатом решения являются значения поправок (4.21) к приближенным значениям корней $\boldsymbol{x}^{(1)}$. Суммы

$$x_i^{(1)} + \delta_i \qquad (i = 1, \ldots, n) \tag{4.24}$$

представляют собой уточненные значения корней исходной системы линейных уравнений (4.3).

Заметим, что при вычислении поправок $\boldsymbol{\delta}$ также накапливается некоторая погрешность. Строго говоря, суммы (4.24) являются не точными решениями исходной системы (4.3), а вторым приближением. Поэтому при необходимости процедуру уточнения корней можно повторить. Разумеется, в качестве оценки точности полученного решения используются относительные погрешности $\left| x_i^{(1)} \right| / \left| \delta_i \right|$.

Современные компьютеры обеспечивают высокую точность вычислений (т.е. малую погрешность округления арифметической операции), но при большом размере исходной системы линейных уравнений (4.1) рекомендуется провести процедуру уточнений корней.

Модификацией метода Гаусса, которая позволяет уменьшать накопление погрешностей в процессе решения, является так называемый ***метод главных элементов***.

Как и в предыдущем параграфе, строится расширенная матрица из n строк и $(n + 1)$ столбцов, т.е. к матрице коэффициентов при неизвестных присоединяется справа столбец свободных членов.

На первом шаге этого метода находится максимальный по модулю коэффициент при неизвестном среди всех элементов расширенной матрицы. Пусть этот элемент располагается на пересечении p-й строки и q-го столбца. Этот элемент a_{pq} называется ***главным элементом***, а строка с номером p называется ***главной строкой***.

Главная строка переставляется на первое место. Из остальных строк вычитается главная, умноженная на отношение a_{iq} / a_{pq} , где i – номер строки ($i = 1, \ldots , n , i \neq p$). В результате этого q-й столбец расширенной матрицы будет состоять из нулей (за исключением главной строки). Это

эквивалентно исключению из уравнений одного из неизвестных. Из расширенной матрицы исключается главная строка и q-й столбец, содержащий нули. При этом получается матрица размером $(n-1)\times n$, т.е. содержащая $(n-1)$ строк и n столбцов.

С новой расширенной матрицей проводится аналогичная операция. Новый главный элемент выбирается среди всех столбцов, кроме самого правого, который состоит из преобразованных свободных членов. Новая главная строка перемещается на второе место вслед за первой главной.

После $(n-1)$-го шага останется расширенная матрица, состоящая из одной строки и двух столбцов. Эта строка вместе с ранее перемещенными главными строками образуют матрицу, которая после необходимой перенумерации неизвестных приобретет вид, аналогичный тому, который получается в обычном методе Гаусса. Матрица преобразованных коэффициентов при неизвестных приобретает треугольный вид. Значения корней вычисляются обратным ходом, идентичным методу Гаусса, описанным в предыдущем параграфе.

Пример 2. Для иллюстрации метода главных элементов решим систему из трех линейных уравнений

$$\begin{aligned} 0{,}34x_1 + 0{,}71x_2 + 0{,}63x_3 &= 2{,}08,\\ 0{,}71x_1 - 0{,}65x_2 - 0{,}18x_3 &= 0{,}17,\\ 1{,}17x_1 - 2{,}35x_2 + 0{,}75x_3 &= 1{,}28. \end{aligned} \qquad (4.25)$$

Соответствующая расширенная матрица имеет вид:

$$\begin{bmatrix} 0{,}34 & 0{,}71 & 0{,}63 & 2{,}08\\ 0{,}71 & -0{,}65 & -0{,}18 & 0{,}17\\ 1{,}17 & -2{,}35 & 0{,}75 & 1{,}28 \end{bmatrix}. \qquad (4.26)$$

Видно, что максимальным по модулю, т.е. главным, является коэффициент a_{32}. Следовательно, главная строка данной системы – третья.

Теперь главную строку (3-ю строку матрицы (4.26)) умножим на отношение $a_{12}/a_{32} = -0{,}3021$ и вычтем ее из 1-й строки матрицы (4.26). Затем главную строку умножим на отношение $a_{22}/a_{32} = 0{,}2766$ и вычтем ее из 2-й строки матрицы (4.26). Вычисления показывают, что вторые элементы в преобразованных строках занулятся. Отбрасывая этот столбец, получаем новую матрицу размером 3×2:

$$\begin{bmatrix} 0{,}6935 & 0{,}8566 & 2{,}4667 \\ 0{,}3864 & -0{,}3875 & -0{,}1840 \end{bmatrix}. \qquad (4.27)$$

В этой матрице требуется выбрать новый ведущий элемент. Это максимальное по модулю число в двух столбцах слева, т.е. $a_{12}^{(1)} = 0{,}8566$. Следовательно, новой (второй) главной строкой является верхняя в матрице (4.27).

Умножим новую главную строку на $a_{22}^{(1)} / a_{12}^{(1)} = -0{,}3875/0{,}8566 = -0{,}4524$ и вычтем ее из второй строки матрицы (4.27). Получаются два ненулевых значения 0,7001 и 0,9319, которые образуют последнюю строку преобразованной матрицы. Соответствующая преобразованная система уравнений примет вид:

$$\begin{aligned} 1{,}17x_1 - 2{,}35x_2 + 0{,}75x_3 &= 1{,}28, \\ 0{,}6935x_1 + 0{,}8566x_3 &= 2{,}4667, \\ 0{,}7001x_1 &= 0{,}9319. \end{aligned} \qquad (4.28)$$

Теперь не составляет труда последовательно вычислить корни данной системы обратным ходом, описанным в предыдущем параграфе:

$$x_1 = 1{,}331; \qquad x_2 = 0{,}693; \qquad x_3 = 1{,}802.$$

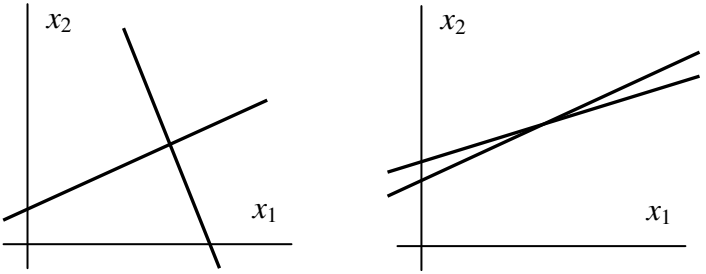
Результаты целесообразно округлить до 0,001 (хотя исходные данные были записаны с точностью до 0,01) и затем провести проверку подстановкой полученных корней в исходную систему уравнений.

Расчет методом главных элементов является несколько более трудоемким по сравнению с обычным методом Гаусса, рассмотренным в §4.3, так как требует на каждом шаге прямого хода совершать поиск главного элемента. Это не является серьезным недостатком метода, так как его реализация проводится на современных компьютерах. Вышеприведенный пример 2 не демонстрирует преимуществ метода главных элементов из-за малого размера решаемой системы. Однако при большом количестве уравнений метод главных элементов позволяет добиться требуемой точности искомых корней без проведения операции уточнения и, следовательно, является более эффективным относительно обычного метода Гаусса.

На точность решения системы линейных уравнений существенно влияют, помимо выше рассмотренных факторов, погрешности исходных данных: коэффициентов матрицы $\boldsymbol{A}$ и свободных членов $\boldsymbol{b}$. Для наглядности рассмотрим геометрические представление системы линейных двух уравнений с двумя неизвестными

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 &= b_1, \\ a_{21}x_1 + a_{22}x_2 &= b_2. \end{aligned} \tag{4.29}$$

Каждое уравнение последней системы можно изобразить прямой линией в системе координат x_1, x_2 (см. рис. 4.1).



а б

Рис. 4.1. Геометрическая схема решения системы двух линейных уравнений с двумя неизвестными.

Корни системы (4.29) на рис. 4.1 представлены координатами точки пересечения прямых линий. Изменение коэффициентов при неизвестных и свободных членов системы (4.29) изобразится сдвигом или поворотом линий. Из рис. 4.1.а видно, что малые изменения величин a_{ij} и b_i ($i, j = 1, 2$) приведут к небольшому сдвигу точки пересечения. Напротив, в случае, изображенном на рис. 4.1.б точка пересечения может переместиться очень сильно даже при малом возмущении чисел a_{ij} и b_i ($i, j = 1, 2$).

Чувствительность решения системы линейных уравнений (4.1) к изменению коэффициентов при неизвестных и свободных членов характеризуется с помощью числа обусловленности $cond(\boldsymbol{A})$.

Для определения этой величины обусловленности сначала требуется ввести понятие нормы вектора и матрицы. В линейной алгебре

используются различные варианты определения норм. В настоящем учебном пособии нормой n-мерного вектора полагается сумма абсолютных значений элементов этого вектора. Например, норма вектора свободных членов системы (4.1) запишется в виде:

$$\| \boldsymbol{b} \| = \sum_{i=1}^{n} |b_i| . \qquad (4.30)$$

За норму матрицы примем максимальное значение из норм векторов, образованных столбцами этой матрицы. Норма матрицы коэффициентов при неизвестных (4.2) выразится так:

$$\| \boldsymbol{A} \| = \max_{j=1,\ldots n} \| \boldsymbol{a}_j \|, \qquad (4.31)$$

где $\boldsymbol{a}_j$ – вектор, состоящий из элементов j-го столбца матрицы $\boldsymbol{A}$.

Числом обусловленности матрицы $\boldsymbol{A}$ называется величина

$$cond(\boldsymbol{A}) = \| \boldsymbol{A} \| \cdot \| \boldsymbol{A}^{-1} \| \qquad (4.32)$$

В теории систем линейных уравнений доказывается, что всегда $cond(\boldsymbol{A}) \geq 1$.

Обозначим $\boldsymbol{\delta b}$ вектор погрешностей свободных членов, $\boldsymbol{\delta x}$ – вектор погрешностей корней. В теории получено следующее соотношение:

$$\frac{\|\delta x\|}{\|x\|} \leq cond(\boldsymbol{A}) \frac{\|\delta b\|}{\|b\|} . \qquad (4.33)$$

Если число обусловленности $cond(\boldsymbol{A})$ много больше единицы, то матрица $\boldsymbol{A}$ называется плохо обусловленной. При этом сравнительно малые относительные погрешности свободных членов могут привести к большим погрешностям корней.

Обозначив $\boldsymbol{\delta A}$ матрицу погрешностей коэффициентов при неизвестных исходной системы линейных уравнений (4.1), можно записать соотношение, которое выражает влияние погрешностей исходных данных на точность корней системы (4.1):

$$\frac{\|\delta x\|}{\|x\|} \le \frac{cond(A)}{1 - cond(A)\dfrac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right). \qquad (4.34)$$

Заметим, что для вычисления числа обусловленности необходимо иметь явный вид обратной матрицы $\boldsymbol{A}^{-1}$. Существуют специальные способы приближенной оценки числа $cond(\boldsymbol{A})$, которые не требуют использования $\boldsymbol{A}^{-1}$. Эти приемы описаны в подробных курсах численных методов, например в [5].

Для решения плохо обусловленных систем линейных уравнений применяются особые методы, изложенные в специальных курсах и учебных пособиях.

§ 4.5. Итерационные методы

Как было указано в начале данной главы, итерационные методы представляют собой методы последовательных приближений. Каждый итерационный метод использует определенное рекуррентное соотношение. По заданному нулевому приближению сначала вычисляется первое, затем по первому второе и т.д. Процесс итераций должен быть построен таким образом, чтобы с ростом числа шагов получаемые приближенные значения корней в принципе сходились к точному решению.

Следует иметь в виду, что исходные данные – коэффициенты при неизвестных и свободные члены в практических задачах часто обладают неустранимой погрешностью. Кроме того, в процессе расчетов на компьютере неизбежны погрешности округления из-за конечной разрядности регистров процессора. В этой связи важным достоинством итерационных методов является то, что можно заранее задать точность искомого решения. Главный же их недостаток – то, что для каждого рекуррентного процесса необходимо исследование условий его сходимости.

Вновь рассмотрим систему линейных уравнений (4.1). Предположим, что все диагональные коэффициенты матрицы $\boldsymbol{A}$ не равны нулю $a_{ii} \neq 0$ (i= 1, 2, ..., n). Разрешим первое уравнение относительно неизвестного x_1, второе уравнение – относительно x_2 и т.д.

В результате получим систему линейных уравнений, эквивалентную данной (4.1).

$$\begin{aligned} x_1 &= d_1 + c_{12}x_2 + c_{13}x_3 + \ldots + c_{1n}x_n, \\ x_2 &= d_2 + c_{21}x_1 + c_{23}x_3 + \ldots + c_{2n}x_n, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n &= d_n + c_{n1}x_1 + c_{n2}x_2 + \ldots + c_{nn-1}x_{n-1}. \end{aligned} \qquad (4.35)$$

В последней системе введены новые обозначения:

$$c_{ij} = -\frac{a_{ij}}{a_{ii}} \quad (\text{при } i \neq j), \quad c_{ij} = 0 \quad (\text{при } i = j), \quad \text{где } i, j = 1, 2, \ldots, n; \qquad (4.36)$$

$$d_i = \frac{b_i}{a_{ii}}, \qquad \text{где } i = 1, 2, \ldots, n. \qquad (4.37)$$

Определим матрицу коэффициентов $\boldsymbol{C}$

$$\boldsymbol{C} = \begin{bmatrix} c_{11} & c_{12} & \ldots & c_{1n} \\ c_{21} & c_{22} & \ldots & c_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ c_{n1} & c_{n2} & \ldots & c_{nn} \end{bmatrix} \qquad (4.38)$$

и вектор-столбец $\boldsymbol{D}$

$$\boldsymbol{D} = \begin{bmatrix} d_1 \\ d_2 \\ \ldots \\ d_n \end{bmatrix}. \qquad (4.39)$$

Перепишем систему (4.35) в матричной форме

$$\boldsymbol{x} = \boldsymbol{D} + \boldsymbol{C}\,\boldsymbol{x}. \qquad (4.40)$$

Решить систему (4.40) можно ***методом простой итерации (методом Якоби)***. Для этого примем вектор $\boldsymbol{D}$ за начальное

приближение вектора $\boldsymbol{x}^{(0)}$ искомых корней. Тогда, подставляя в правую часть уравнений (4.40) вектор $\boldsymbol{x}^{(0)}$, получим в левой части вектор-столбец первого приближения корней:

$$\boldsymbol{x}^{(1)} = \boldsymbol{D} + \boldsymbol{C}\,\boldsymbol{x}^{(0)}.$$

Последовательно повторяя эту операцию, вычисляем (k+1)-е приближения с помощью следующего рекуррентного матричного соотношения:

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{D} + \boldsymbol{C}\,\boldsymbol{x}^{(k)}. \qquad (4.41)$$

Распишем формулы последовательных приближений в развернутом виде:

$$x_i^{(0)} = d_i, \qquad x_i^{(k+1)} = d_i + \sum_{j=1}^{n} c_{ij} x_j^{(k)}\ , \qquad (4.42)$$

где $i = 1, 2, ..., n$; $k = 0, 1, 2, ...$.

Метод простой итерации основан на том, что последовательность векторов-столбцов $\boldsymbol{x}^{(k)}$ (k = 0, 1, 2, ...) сходится к точному решению исходной системы (4.1) при определенных условиях, которые будут сформулированы ниже.

Достаточное условие сходимости итерационного процесса (4.41) может быть выражено в виде:

$$\| \boldsymbol{C} \| < 1. \qquad (4.43)$$

В курсе линейной алгебры доказывается, что при условии (4.43) итерационный процесс (4.41) сходится к единственному решению исходной системы (4.1) при любом начальном приближении $\boldsymbol{x}^{(0)}$. Следовательно, в принципе, не обязательно за исходное приближение $\boldsymbol{x}^{(0)}$ брать столбец свободных членов (4.39).

Однако практика показывает, что из-за погрешностей исходных данных и округления при выполнении операций процессором сходимость процесса к точному решению может быть нарушена.

Для использования на практике метода простой итерации, прежде всего, необходимо иметь все диагональные коэффициенты матрицы $\boldsymbol{A}$ не равными нулю $a_{ii} \neq 0$ (i = 1, 2, ..., n). Для невырожденной матрицы $\boldsymbol{A}$

это может быть выполнено перестановкой уравнений исходной системы (4.1).

Из определения нормы матрицы следует, что условие (4.43) у матрицы $\boldsymbol{C}$ выполнится, если ее элементы малы по абсолютной величине. Это достигается, если диагональные элементы a_{ii} достаточно велики по модулю относительно остальных элементов a_{ij} ($i \neq j$) матрицы $\boldsymbol{A}$, что следует из определения (4.36). Можно доказать, что условие сходимости (4.43) будет выполнено, если справедливы n неравенств:

$$| a_{ii} | > \sum_{\substack{j=1 \\ j \neq i}}^{n} \left| a_{ij} \right|, \quad i = 1, 2, ..., n \, . \tag{4.44}$$

Иначе говоря, метод простой итерации даст единственное решение системы линейных уравнений (4.1), если в каждом уравнении модуль диагонального коэффициента превышает сумму модулей остальных коэффициентов в соответствующей строке исходной системы, т.е. сумму модулей остальных коэффициентов при неизвестных в этом уравнении.

Выполнить требование (4.44) можно, применяя линейные преобразования уравнений исходной системы (4.1).

Пример 3. Дана следующая система из 4-х линейных уравнений:

$$\begin{aligned}
&2x_1 + 3x_2 - 4x_3 + x_4 = 3, && \text{(а)} \\
&x_1 - 2x_2 - 5x_3 + x_4 = 2, && \text{(б)} \\
&5x_1 - 3x_2 + x_3 - 4x_4 = 1, && \text{(в)} \\
&10x_1 + 2x_2 - x_3 + 2x_4 = -4. && \text{(г)}
\end{aligned}$$

Если взять уравнения (б) и (г) в качестве 3-го и 1-го уравнений новой системы, для них будет выполняться условие (4.44). Разность уравнений (а) и (б) дает 2-е уравнение. Наконец линейная комбинация уравнений 2((а) – (б) + (в)) – (г) даст 4-е уравнение преобразованной системы. В результате получается система

$$\begin{aligned}
&10x_1 + 2x_2 - x_3 + 2x_4 = -4, \\
&x_1 + 5x_2 + x_3 + 0 = 1, \\
&x_1 - 2x_2 - 5x_3 + x_4 = 2, \\
&3x_1 + 0x_2 + 0x_3 - 9x_4 = 10.
\end{aligned}$$

Для всех уравнений последней системы выполнены условия (4.44). Можно эту систему преобразовывать к виду (4.35) и решать методом простой итерации (4.41).

В качестве иллюстрации скорости сходимости итерационного процесса (4.41) еще один пример.

Пример 4. Пусть система состоит из трех следующих линейных уравнений.

$$\begin{aligned} 4\,x_1 + 0{,}24\,x_2 - 0{,}08\,x_3 &= 8,\\ 0{,}09\,x_1 + 3\,x_2 - 0{,}15\,x_3 &= 9,\\ 0{,}04\,x_1 - 0{,}08\,x_2 + 4\,x_3 &= 20. \end{aligned} \qquad (4.45)$$

Видно, что для данной системы условие (4.44) выполнено и, следовательно, можно для решения применить метод простой итерации.

Сначала преобразуем систему (4.45) к стандартному виду (4.35):

$$\begin{aligned} x_1 &= 2 - 0{,}06\,x_2 + 0{,}02\,x_3,\\ x_2 &= 3 - 0{,}03\,x_1 + 0{,}05\,x_3,\\ x_3 &= 5 - 0{,}01\,x_1 + 0{,}02\,x_2. \end{aligned} \qquad (4.46)$$

В качестве начального приближения возьмем вектор свободных членов системы уравнений (4.46): $x_1^{(0)} = 2$; $x_2^{(0)} = 3$; $x_3^{(0)} = 5$. Подставляя эти числа в правые части уравнений (4.46), получим в левых частях первое приближение решения. Результаты нескольких итераций сведены в следующую таблицу 4.1.

Значения, приведенные в таблице 4.1, показывают, что решение системы линейных уравнений (4.45) с точностью до пяти знаков после запятой достигается уже после 4-й итерации. Последующие итерации не изменяют значения указанных десятичных разрядов. Если требуется вычислить корни с точностью до 0,001, то достаточно провести лишь 3 итерации. Полученное решение будет достаточно точным в рамках заданной погрешности.

Таблица 4.1

Номер итерации	x_1	x_2	x_3
0	2	3	5
1	1,92	3,19	5,04
2	1,9094	3,1944	5,0446
3	1,909228	3,194948	5,044794
4	1,909199	3,194963	5,044807
5	1,909198	3,194964	5,044807

Пример 4 демонстрирует, что после преобразования исходной системы к виду (4.35) остается проводить операции умножения и сложения, вычисляя правые части этой системы. Хранить в оперативной памяти компьютера приходится только неизменяемую матрицу $\boldsymbol{C}$.

Если изначально задана требуемая абсолютная погрешность ε вычисления корней, то на каждом шаге итераций проверяются условия

$$\left| x_i^{(k)} - x_i^{(k-1)} \right| < \varepsilon, \;\; i = 1, \ldots, n. \qquad (4.47)$$

где k – номер итерации, i – номер уравнения. Если хотя бы одно из условий (4.47) не выполняется, то проводится следующая итерация. В результате все корни решаемой системы уравнений будут вычислены с абсолютной погрешностью $\leq \varepsilon$.

При решении прикладных задач более часто используется другой итерационный метод, называемый ***методом Зейделя***.

Метод Зейделя базируется на вышеописанной схеме и использует рекуррентное соотношение (4.41). Особенностью этого метода является то, что при вычислении k-го приближения корня x_i системы линейных уравнений (4.1) учитываются уже вычисленные ранее k-е приближения корней $x_1, x_2, \ldots, x_{i-1}$.

Пусть исходная система (4.1) преобразована в приведенную систему (4.35). Выбирается вектор начального приближения искомых корней. Итерации проводятся по схеме (4.41) с одной оговоркой. Если уже вычислено k-е приближение корней $\boldsymbol{x}^{(k)}$, то следующее $(k+1)$-е приближение $\boldsymbol{x}^{(k+1)}$ строится по следующим формулам:

$$
\begin{aligned}
x_1^{(k+1)} &= d_1 + \sum_{j=1}^{n} c_{1j} x_j^{(k)} , \\
x_2^{(k+1)} &= d_2 + c_{21} x_1^{(k+1)} + \sum_{j=2}^{n} c_{2j} x_j^{(k)} , \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
x_i^{(k+1)} &= d_i + \sum_{j=1}^{i-1} c_{ij} x_j^{(k+1)} + \sum_{j=i}^{n} c_{ij} x_j^{(k)} , \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
x_n^{(k+1)} &= d_n + \sum_{j=1}^{n-1} c_{ij} x_j^{(k+1)} + c_{nn} x_n^{(k)} .
\end{aligned}
\qquad (4.48)
$$

Как и в предыдущем методе, после нескольких итераций по данной схеме корни $\boldsymbol{x}^{(k+1)}$ достигают предельных значений (конечно, в пределах определенного количества десятичных разрядов). Следовательно, использование рекуррентного процесса (4.48) позволяет за конечное число итераций получить значения корней исходной системы (4.1) с заданной погрешностью.

Метод Зейделя иногда дает более быструю сходимость по сравнению с методом простой итерации. Условия сходимости этих методов, вообще говоря, не совпадают. Существуют такие матрицы коэффициентов при неизвестных, для которых метод Зейделя сходится, а метод простой итерации – нет. Встречаются и противоположные ситуации.

Для формулировки условий сходимости метода Зейделя целесообразно ввести понятие нормальной системы линейных уравнений.

Система линейных уравнений вида (4.1) называется ***нормальной***, если выполняются два следующих условия.

1. Матрица $\boldsymbol{A}$ коэффициентов при неизвестных симметрична, т.е. $a_{ij}=a_{ji}$.

2. Квадратичная форма U, соответствующая матрице $\boldsymbol{A}$,

$$
U = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j \qquad (4.49)
$$

положительно определена (т.е. принимает при любых значениях аргументов x_i неотрицательные значения, причем обращается в нуль, только когда все аргументы x_i равны нулю).

Если матрица коэффициентов $\boldsymbol{A} = [\ a_{ij}\ ]$ неособенная (т.е. $\det \boldsymbol{A} \neq 0$), то система линейных уравнений общего вида (4.1) может быть приведена к нормальному виду. Для этого обе части матричного уравнения (4.3) надо умножить слева на транспонированную матрицу $\boldsymbol{A}' = [\ a_{ji}\ ]$:

$$\boldsymbol{A}'\,\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{A}'\,\boldsymbol{b}. \tag{4.50}$$

Произведение квадратных матриц $\boldsymbol{A}'\boldsymbol{A}$ представляет собой тоже матрицу $\boldsymbol{A}_N$ размером $n \times n$. В курсе линейной алгебры доказывается, что матрица $\boldsymbol{A}_N$ симметрическая и соответствует положительно определенной квадратичной форме. Произведение $\boldsymbol{A}'\boldsymbol{b} = \boldsymbol{b}_N$ является n-мерным вектором. Таким образом, нормализация исходной системы линейных уравнений (4.1) дает эквивалентную систему вида

$$\boldsymbol{A}_N\,\boldsymbol{x} = \boldsymbol{b}_N. \tag{4.51}$$

После получения нормальной системы (4.51) можно ее преобразовать к виду (4.35). В линейной алгебре доказана следующая теорема: если система линейных уравнений вида (4.1) – нормальная, то процесс Зейделя для эквивалентной системы (4.35) сходится к точному решению системы при любом выборе начального приближения.

Пример 5. Рассмотрим следующую систему трех линейных уравнений:

$$\begin{aligned} 9\,x_1 - 2\,x_2 + x_3 &= 8, \\ 2\,x_1 - 7\,x_2 + x_3 &= -4, \\ x_1 + 3\,x_2 + 8\,x_3 &= 12. \end{aligned} \tag{4.52}$$

Легко видеть, что для уравнений системы (4.52) выполнены условия (4.44), поэтому возможно решить данную систему итерационным методом. Используем метод Зейделя.

Приведем систему (4.52) к виду (4.35), необходимому для проведения итераций:

$$
\begin{aligned}
x_1 &= 0{,}88889 + 0{,}22222\, x_2 - 0{,}11111\, x_3,\\
x_2 &= 0{,}57142 + 0{,}28571\, x_1 + 0{,}14286\, x_3,\\
x_3 &= 1{,}5 - 0{,}125\, x_1 - 0{,}375\, x_2.
\end{aligned}
\qquad (4.53)
$$

Так как выполнение условий (4.44) обеспечивает сходимость при любом начальном приближении, в качестве такового выберем нулевое: $x_1^{(0)}=0;\ x_2^{(0)}=0;\ x_3^{(0)}=0$. Подставим эти значения в правую часть первого уравнения системы (4.53) и легко получим первое приближение первого корня $x_1^{(1)}=0{,}88889$. При вычислении первого приближения второго корня $x_2^{(1)}$ используем уже полученное значение $x_1^{(1)}$:

$$x_2^{(1)} = 0{,}57142 + 0{,}28571\ x_1^{(1)} + 0{,}14286\ x_3^{(0)} = 0{,}82539.$$

Аналогично вычислим первое приближение третьего корня

$$x_3^{(1)} = 1{,}5 - 0{,}125\ x_1^{(1)} - 0{,}375\ x_2^{(1)} = 1{,}07937.$$

Продолжим итерации методом Зейделя, а результаты запишем в таблицу 4.2.

В том, что точным решением системы (4.52) является $x_1 = x_2 = x_3 = 1$, легко убедиться подстановкой полученных значений в исходную систему.

Таблица 4.2

Номер итерации	x_1	x_2	x_3
0	0	0	0
1	0,88889	0,82539	1,07937
2	0,95238	0,99773	1,0068
3	0,99874	1,00061	0,99993
4	1,00014	1,00003	0,99997
5	1,00001	0,999999	0,999999
6	1,000000	1,000000	1,000000

Из таблицы 4.2 видно, что значения всех трех корней быстро сходятся к единице. Точность до шестого знака после запятой достигается на шестой итерации.

Глава 5. ВЫЧИСЛЕНИЕ ДЕТЕРМИНАНТОВ

Каждая квадратная матрица $\boldsymbol{A}$ характеризуется определенным числом $\det \boldsymbol{A}$, которое называется ***детерминантом*** или ***определителем*** этой матрицы. Порядок детерминанта совпадает с порядком соответствующей матрицы.

Как следует из предыдущей главы, при решении системы линейных уравнений полезно (или даже необходимо) знать численное значение детерминанта матрицы коэффициентов при неизвестных. Значения детерминантов матриц требуются и в других прикладных задачах физики. Поэтому целесообразно в этой главе рассмотреть вопрос о вычислении детерминантов.

В линейной алгебре величина детерминанта определяется индуктивно. Квадратная матрица порядка $n = 1$ состоит из единственного элемента a_{11}. Детерминант такой матрицы принимается равным этому элементу. Детерминант квадратной матрицы порядка $n = 2$ определяется следующей формулой:

$$\det \boldsymbol{A} = a_{11} \cdot a_{22} - a_{12} \cdot a_{21}. \tag{5.1}$$

Детерминанты порядка $n > 2$ определяются через детерминанты более низких порядков с помощью разложения по минорам. Как известно из линейной алгебры, каждому элементу a_{ij} квадратной матрицы $\boldsymbol{A}$ порядка n ставится в соответствие минор M_{ij}. Минор M_{ij} представляет собой детерминант матрицы $\mathbf{A}_{ij}$, которая получается из исходной $\boldsymbol{A}$ вычеркиванием i-й строки и j-го столбца. Ясно, что каждая матрица $\mathbf{A}_{ij}$ и соответствующий минор M_{ij} имеют порядок $n - 1$.

В линейной алгебре доказана теорема, что детерминант n-го порядка может быть вычислен по следующей формуле:

$$\det \boldsymbol{A} = \sum_{j=1}^{n} (-1)^{i+j} \, a_{ij} \cdot M_{ij}, \tag{5.2}$$

где номер строки может принимать любое значение $i = 1, \ldots, n$.

Формула (5.2) называется разложением детерминанта по минорам i-й строки. Справедлива и другая формула вычисления детерминанта n-го порядка разложением по минорам j-го столбца:

$$\det \boldsymbol{A} = \sum_{i=1}^{n} (-1)^{i+j}\, a_{ij} \cdot M_{ij}, \qquad (5.3)$$

где номер j может быть выбран любым из множества: 1, …, n.

Таким образом, для вычисления одного детерминанта n-го порядка необходимо вычислить n детерминантов (n – 1)-го порядка. Для вычисления каждого минора можно вновь применить формулу (5.2). В результате, чтобы вычислить один детерминант n-го порядка этим методом приходится рассчитать $n!/2$ детерминантов 2-го порядка. Кроме того, необходимо выполнить умножения миноров на коэффициенты a_{ij} исходной матрицы и провести суммирования по формулам (5.2) или (5.3). Ясно, что для высокого порядка вычисление детерминантов разложением по минорам является длительной операцией даже при использовании персональных компьютеров.

В линейной алгебре также используется метод выражения детерминантов непосредственно через коэффициенты матрицы a_{ij} без промежуточного вычисления миноров. Этот метод можно представить состоящим из следующих этапов.

Вначале составляется произведение коэффициентов a_{ij} вида $a_{1\alpha}a_{2\beta}...a_{n\omega}$, где индексы α, β, ..., ω взяты из множества целых чисел (1, 2, ..., n) и все различны. В качестве первого набора можно взять $\alpha = 1$, $\beta=2$, ..., $\omega = n$. Затем перебираются всевозможные перестановки индексов α, β, ..., ω в произведениях $a_{1\alpha}a_{2\beta}...a_{n\omega}$, начиная с исходной. Если перестановка нечетная, то произведение $a_{1\alpha}a_{2\beta}...a_{n\omega}$ умножается на -1. После перебора всех перестановок индексов α, β, ..., ω все полученные произведения суммируются. Заметим, что количество перестановок равно $n!$.

Справедливость такого представления для детерминанта 2-го порядка видна непосредственно из формулы (5.1). Детерминант 3-го порядка, построенный последним методом, дает следующее выражение:

$$\begin{aligned}\det \boldsymbol{A} = {} & a_{11}a_{22}a_{33} - a_{12}a_{21}a_{33} + a_{13}a_{21}a_{32} - a_{11}a_{23}a_{32} + \\ & + a_{12}a_{23}a_{31} - a_{13}a_{22}a_{31}, \end{aligned} \qquad (5.4)$$

что совпадает с результатом, полученным по методу разложения по минорам.

Пример 1. Необходимо вычислить детерминант следующей матрицы размера 4×4:

$$\det \boldsymbol{A} = \begin{bmatrix} 5 & -4 & 0 & 2 \\ -1 & 1 & 1 & -1 \\ 2 & 3 & 3 & -6 \\ 1 & 0 & 2 & -1 \end{bmatrix}. \qquad (5.5)$$

Разложим данный детерминант 4-го порядка на миноры по первой строке. Применив формулу (5.2) и пользуясь тем, что элемент $a_{13} = 0$, получим разложение:

$$\det \boldsymbol{A} = 5\begin{vmatrix} 1 & 1 & -1 \\ 3 & 3 & -6 \\ 0 & 2 & -1 \end{vmatrix} - (-4)\begin{vmatrix} -1 & 1 & -1 \\ 2 & 3 & -6 \\ 1 & 2 & -1 \end{vmatrix} - 2\begin{vmatrix} -1 & 1 & 1 \\ 2 & 3 & 3 \\ 1 & 0 & 2 \end{vmatrix}. \qquad (5.6)$$

Далее перед нами два разных пути. Можно каждый из трех полученных детерминантов 3-го порядка вычислить по формуле (5.4). Иначе можно детерминанты 3-го порядка предварительно разложить и свести к вычислению нескольких детерминантов 2-го порядка. Если проводить разложение по минорам нижних строк, то (из-за того, что $a_{42} = 0$) для вычисления первого и третьего детерминантов в (5.6) потребуется вычислить только по два детерминанта 2-го порядка. Предлагаем читателям провести необходимые выкладки и убедиться, что $\det \boldsymbol{A} = -6$.

Описанный алгоритм может быть запрограммирован и реализован на персональном компьютере. Нетрудно подсчитать, что для вычисления детерминанта n-го порядка необходимо выполнить $n \cdot n!$ умножений и $n!$ сложений (или вычитаний). Следовательно, метод разложения по минорам практически не пригоден для вычисления детерминантов высоких порядков.

Более эффективный метод (требующий гораздо меньшего количества операций) базируется на следующей теореме линейной алгебры. *Если к элементам любой строки (или любого столбца) детерминанта прибавить соответствующие элементы другой строки (или столбца), умноженные на произвольный множитель, то величина детерминанта не изменится*. В частности, подобными операциями исходная квадратная матрица приводится к треугольному виду. Способ

приведения квадратной матрицы к треугольному виду аналогичен процедуре прямого хода в методе Гаусса при решении системы линейных уравнений.

Процедура начинается с выбора ведущего элемента. Это может быть любой коэффициент a_{i1}, не равный нулю. Переставляя строки (или столбцы) детерминанта, всегда можно ведущим элементом сделать коэффициент a_{11}. При этом следует помнить, что при перестановке двух строк (или двух столбцов) знак детерминанта изменяется на противоположный.

Затем коэффициент a_{11} выносится из первого столбца в качестве общего множителя детерминанта. Вычитаем из элементов a_{ij} каждого j-го столбца ($j \geq 2$) соответствующие элементы a_{i1} первого столбца, умноженные на a_{1j}. Преобразование элементов можно записать в виде:

$$a_{ij}^{(1)} = a_{ij} - a_{i1} \frac{a_{1j}}{a_{11}}, \quad i = 1, ..., n, \quad j = 2, ..., n. \tag{5.7}$$

В преобразованном детерминанте элемент $a_{11}^{(1)} = 1$. Остальные элементы первой строки равны нулю. Разложение этого детерминанта по элементам первой строки, согласно (5.2), показывает, что он равен детерминанту порядка $(n - 1)$, составленному из элементов $a_{ij}^{(1)}$ ($i, j = 2, ..., n$). Таким образом, исходный детерминант n-го порядка выразится в виде произведения:

$$\det \boldsymbol{A} = a_{11} \Delta_{n-1}, \tag{5.8}$$

где

$$\Delta_{n-1} = \begin{vmatrix} a_{22}^{(1)} & a_{23}^{(1)} & ... & a_{2n}^{(1)} \\ a_{32}^{(1)} & a_{33}^{(1)} & ... & a_{3n}^{(1)} \\ ... & ... & ... & ... \\ a_{n2}^{(1)} & a_{n3}^{(1)} & ... & a_{nn}^{(1)} \end{vmatrix} \tag{5.9}$$

новый детерминант $(n - 1)$-го порядка.

Следующий шаг заключается в применении к новому детерминанту Δ_{n-1} той же процедуры. Выбирается новый ведущий элемент, не равный нулю. Переставляя при необходимости строки или столбцы нового

детерминанта, делаем ведущим элемент $a_{22}^{(1)}$. Выносим его из первого столбца нового детерминанта в качестве общего множителя. Процедура продолжается, причем на каждом шаге данного алгоритма уменьшается порядок детерминанта. Преобразования элементов приводят к тому, что первая строка каждого нового детерминанта начинается с единицы. Остальные элементы этой строки – нули.

В результате, если на каждом шаге удается найти ненулевой ведущий элемент, искомый детерминант n-го порядка представится в виде произведения этих ведущих элементов:

$$\det \boldsymbol{A} = a_{11} \cdot a_{22}^{(1)} \cdot \ldots \cdot a_{nn}^{(n-1)}. \qquad (5.10)$$

Этот способ вычисления детерминанта называется ***методом Гаусса***.

Если на каком-либо шаге не удается отыскать ненулевой ведущий элемент, то это значит, что детерминант исходной матрицы равен нулю.

В терминологии матричной алгебры можно сказать, что проведенное преобразование исходной квадратной матрицы к треугольной позволяет выразить ее детерминант в виде произведения диагональных элементов полученной треугольной матрицы.

Можно подсчитать, что количество арифметических операций для вычисления детерминанта n-го порядка методом Гаусса не превышает n^3. Следовательно, метод Гаусса предпочтительнее рассмотренных выше уже при $n > 5$.

Пример 2. Вычислим детерминант (5.5), приведенный в примере 1 данной главы, но на этот раз используем метод Гаусса, изложенный выше.

Первый ведущий элемент $a_{11} = 5$. Преобразуем 2-ю строку данной матрицы по формуле (5.7):

$$a_{2j}^{(1)} = a_{2j} - a_{21} \frac{a_{1j}}{a_{11}}, \quad j = 2, 3, 4 \text{ (номер столбца)}.$$

$$a_{22}^{(1)} = 1-(-1)\cdot(-4)/5 = 1/5, \quad a_{23}^{(1)} = 1-(-1)\cdot 0/5 = 1, \quad a_{24}^{(1)} = -1-(-1)\,2/5 = -3/5.$$

Аналогично преобразуются 3-я и 4-я строки данной матрицы:

$$a_{3j}^{(1)} = a_{3j} - a_{31}\frac{a_{1j}}{a_{11}}, \quad a_{4j}^{(1)} = a_{4j} - a_{41}\frac{a_{1j}}{a_{11}}, \quad j = 2, 3, 4.$$

Получаем новый детерминант

$$\Delta_3 = \begin{vmatrix} a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} \\ a_{32}^{(1)} & a_{33}^{(1)} & a_{34}^{(1)} \\ a_{42}^{(1)} & a_{43}^{(1)} & a_{44}^{(1)} \end{vmatrix} = \begin{vmatrix} 1/5 & 1 & -3/5 \\ 23/5 & 3 & -34/5 \\ 4/5 & 2 & -7/5 \end{vmatrix}.$$

Новый ведущий элемент $a_{22}^{(1)} = 1/5$. Предпоследняя и последняя строки детерминанта Δ_3 преобразуются по тому же алгоритму (см. первую формулу в (4.15)):

$$a_{3j}^{(2)} = a_{3j}^{(1)} - a_{32}^{(1)}\frac{a_{2j}^{(1)}}{a_{22}^{(1)}}, \quad a_{4j}^{(2)} = a_{4j}^{(1)} - a_{42}^{(1)}\frac{a_{2j}^{(1)}}{a_{22}^{(1)}}, \quad j = 3, 4.$$

Следующий детерминант получает вид:

$$\Delta_2 = \begin{vmatrix} a_{33}^{(2)} & a_{34}^{(2)} \\ a_{43}^{(2)} & a_{44}^{(2)} \end{vmatrix} = \begin{vmatrix} -20 & 7 \\ -2 & 1 \end{vmatrix}.$$

Теперь вычислим последний сомножитель произведения (5.10):

$$a_{nn}^{(n-1)} = a_{44}^{(3)} = a_{44}^{(2)} - a_{43}^{(2)}\frac{a_{34}^{(2)}}{a_{33}^{(3)}} = 1-(-2)\cdot 7/(-20) = 3/10.$$

Искомый детерминант равен $\det \boldsymbol{A} = a_{11}\, a_{22}^{(1)}\, a_{33}^{(2)}\, a_{44}^{(3)} = 5\cdot\frac{1}{5}\cdot(-20)\cdot\frac{3}{10} = -6.$

Конечно, детерминант Δ_2 можно было бы сосчитать по формуле (5.1) и представить результат в виде произведения $\det \boldsymbol{A} = a_{11}\cdot a_{22}^{(1)}\cdot\Delta_2$. Искомое значение $\det \boldsymbol{A} = -6$ достигается еще быстрее.

Глава 6. ОБРАЩЕНИЕ МАТРИЦ

Как известно из линейной алгебры, обратной матрицей $\boldsymbol{A^{-1}}$ по отношению к данной $\boldsymbol{A}$ называется матрица, которая будучи умноженной на $\boldsymbol{A}$ как справа, так и слева, дает единичную матрицу $\boldsymbol{E}$:

$$\boldsymbol{A}\cdot\boldsymbol{A^{-1}} = \boldsymbol{A^{-1}}\cdot\boldsymbol{A} = \boldsymbol{E}. \tag{6.1}$$

Единичную матрицу $\boldsymbol{E}$ удобно кратко записывать с помощью символов Кронекера:

$$\boldsymbol{E} = [\ \delta_{ij}\ ], \tag{6.2}$$

где

$$\delta_{ij} = \begin{cases} 1 & (i = j) \\ 0 & (i \neq j) \end{cases}. \tag{6.3}$$

Матрицы, обратные данным, используются во многих прикладных задачах, в т.ч. и физических. Следовательно, существует потребность в методах, позволяющих вычислять элементы обратной матрицы, выражая их через элементы исходной.

В курсе линейной алгебры доказано, что обратная матрица $\boldsymbol{A^{-1}}$ существует, если исходная матрица $\boldsymbol{A}$ является неособенной, т.е. ее детерминант не равен нулю ($\det \boldsymbol{A} \neq 0$). Методы вычисления детерминантов приведены в предыдущей главе.

Одна из теорем линейной алгебры утверждает, что обратная матрица $\boldsymbol{A^{-1}}$ может быть представлена как транспонированная матрица алгебраических дополнений исходной $\boldsymbol{A}$, причем каждый элемент этой транспонированной матрицы должен быть поделен на детерминант $\det \boldsymbol{A}$. Алгебраическое дополнение A_{ij} любого элемента матрицы a_{ij} выражается через его минор M_{ij}:

$$A_{ij} = (-1)^{i+j} \cdot M_{ij}. \tag{6.4}$$

Таким образом, обратная матрица $\boldsymbol{A^{-1}}$ выражается в следующем виде:

$$\boldsymbol{A}^{-1} = \begin{bmatrix} A_{11}/\Delta & A_{21}/\Delta & \ldots & A_{n1}/\Delta \\ A_{12}/\Delta & A_{22}/\Delta & \ldots & A_{n2}/\Delta \\ \ldots & \ldots & \ldots & \ldots \\ A_{1n}/\Delta & A_{1n}/\Delta & \ldots & A_{nn}/\Delta \end{bmatrix}, \qquad (6.5)$$

где использовано обозначение $\Delta = \det \boldsymbol{A}$.

В принципе обратную матрицу можно вычислить по формуле (6.5). Операция транспонирования не представляет трудности. Но для вычисления каждого алгебраического дополнения необходимо вычислить минор – детерминант порядка $(n - 1)$. Следовательно, для расчета обратной матрицы размером $(n \times n)$ необходимо вычислить один детерминант n-го порядка и n^2 детерминантов $(n-1)$-го порядка. При больших значениях n вычисление элементов обратной матрицы по формуле (6.5) требует значительного времени, даже при использовании современных персональных компьютеров, поэтому описанный метод обращения матриц практически не используется.

За годы развития прикладной математики было разработано много различных способов обращения матриц: разбиение на клетки, окаймление, представление произведением треугольных матриц (LU-факторизация), схема Халецкого, и т.д. Эффективность некоторых методов повышается при специфическом виде обращаемой матрицы. Для плохо обусловленных матриц, т.е. тех, у которых число обусловленности (4.32) много больше единицы, разработаны специальные методы обращения, обеспечивающие достаточную точность вычислений.

В эпоху распространения персональных компьютеров одним из широко используемых методов обращения хорошо обусловленных матриц является ***метод Гаусса***.

Пусть нам дана хорошо обусловленная квадратная матрица $\boldsymbol{A}$ порядка n вида (4.2). Обозначим через x_{ij} элементы обратной матрицы $\boldsymbol{A}^{-1}$. Умножим исходную матрицу $\boldsymbol{A}$ на обратную $\boldsymbol{A}^{-1}$, используя свойство (6.1). Результат перемножения этих матриц можно записать в следующем общем виде:

$$\sum_{k=1}^{n} a_{ik} x_{kj} = \delta_{ij} \qquad (i, j = 1, \ldots, n), \qquad (6.6)$$

где δ_{ij} – символ Кронекера (6.3).

Система (6.6) содержит n^2 уравнений и является линейной относительно искомых неизвестных x_{ij}. Нетрудно заметить, что совокупность уравнений (6.6) состоит из n линейных систем, которые имеют одну и ту же матрицу коэффициентов при неизвестных (4.2). Друг от друга системы уравнений отличаются векторами свободных членов δ_{ij}, причем в каждой системе только один свободный член равен единице (остальные равны нулю). Каждой j-й набор свободных членов определяет j-й столбец x_{ij} ($i = 1, \ldots, n$) искомой обратной матрицы $\boldsymbol{A}^{-1}$.

Поиск элементов x_{ij} обратной матрицы $\boldsymbol{A}^{-1}$ решением системы уравнений (6.6) эффективно осуществить алгоритмом Гаусса, который ранее успешно применялся в главах 4 и 5. Наличие нескольких столбцов свободных членов при неизменной матрице коэффициентов при неизвестных в системе уравнений (6.6) обусловливает некоторую специфику применения метода Гаусса к решению данной задачи обращения матрицы.

Прежде всего, строится расширенная матрица размером $n \times 2n$. К исходной матрице $\boldsymbol{A}$ приписывается справа еще единичная матрица $\boldsymbol{E}$ того же порядка.

Далее к расширенной матрице $(n - 1)$ раз применяется рекуррентное преобразование, аналогичное прямому ходу метода Гаусса в ходе решения системы линейных уравнений (см. § 4.3 и соотношения (4.17)). В результате левая половина расширенной матрицы приводится к диагональному виду.

Затем реализуется процедура обратного хода метода Гаусса, причем на каждом шаге вычисления проводятся для всех столбцов свободных членов, что дает очередную строку искомой обратной матрицы.

В результате этих действий левая половина расширенной матрицы превратится в единичную, а в правой половине сформируется матрица $\boldsymbol{A}^{-1}$ размером $n \times n$, обратная исходной $\boldsymbol{A}$. В справедливости того, что правая половина преобразованной матрицы является обратной данной, можно убедиться непосредственным перемножением. Произведение полученной квадратной матрицы и исходной $\boldsymbol{A}$ дает единичную матрицу $\boldsymbol{E}$ размером $n \times n$.

Пример 1. Используя метод Гаусса, найдем матрицу, обратную данной:

$$\boldsymbol{A} = \begin{bmatrix} 2 & -1 & 1 & 2 \\ 1 & 2 & -1 & 1 \\ 3 & 0 & -1 & -3 \\ 1 & -1 & 1 & 3 \end{bmatrix}. \qquad (6.7)$$

Сначала составим расширенную матрицу вышеописанного типа и поместим ее в таблицу 6.1. В крайнем левом столбце запишем порядковые номера строк. В следующие 4 столбца поместим исходную матрицу, в четырех правых столбцах – единичную матрицу размером 4 × 4. Элементы расширенной матрицы обозначим c_{ij}, где i – номер строки, j – номер столбца.

Верхний левый элемент выбираем в качестве ведущего. Выделим его в таблице прямоугольником. Под штриховой линей в таблице 6.1 запишем строку, получаемую делением первой строки исходной расширенной матрицы на ведущий элемент. Присвоим полученной строке номер k=1. Очевидно, что первый элемент этой строки равен единице.

Таблица 6.1

i	j=1	j=2	j=3	j =4	j=5	j=6	j=7	j=8
1	2	−1	1	2	1	0	0	0
2	1	2	−1	1	0	1	0	0
3	3	0	−1	−3	0	0	1	0
4	1	−1	1	3	0	0	0	1
k=1	1	−1/2	1/2	1	1/2	0	0	0

Затем строки с номерами i = 2, 3 и 4 исходной расширенной матрицы подвергаем преобразованию Гаусса, т.е. умножаем нижнюю строку с номером k=1 на коэффициенты c_{i1} (i = 2, 3, 4) и вычитаем из соответствующей i-й строки. Иначе говоря, вычисляем элементы новой матрицы по формулам (4.17). Получаем новую матрицу из трех строк, у которой первый столбец с j=1 состоит из нулей. Запишем полученную матрицу в таблицу 6.2.

Теперь ведущим берем элемент c_{22} = 5/2. Его также выделяем прямоугольником и делим на него строку с номером i = 2 в таблице 6.2. Результат деления записываем ниже штриховой линии в таблице 6.2 в

строке, обозначенной k=2. Затем умножим полученную строку с номером k = 2 на коэффициенты c_{32} и c_{42} таблицы 6.2 и вычтем из соответствующих двух строк с номерами i = 3 и i = 4 той же таблицы.

Таблица 6.2

i	j=1	j=2	j=3	j =4	j=5	j=6	j=7	j=8
2	0	5/2	–3/2	0	–1/2	1	0	0
3	0	3/2	–5/2	–6	–3/2	0	1	0
4	0	–1/2	1/2	2	–1/2	0	0	1
k=2	0	1	–3/5	0	–1/5	2/5	0	0

Получается матрица из двух строк, у которой уже два левых столбца состоят из нулей. Эту новую матрицу запишем в таблицу 6.3.

Таблица 6.3

i	j=1	j=2	j=3	j =4	j=5	j=6	j=7	j=8
3	0	0	–8/5	–6	–6/5	–3/5	1	0
4	0	0	1/5	2	–3/5	1/5	0	1
k=3	0	0	1	15/4	3/4	3/8	–5/8	0

На следующем шаге ведущим элементом является c_{33} = –8/5. На него делится строка с номером i = 3 в таблице 6.3, а результат записываем под штриховой линией в таблице 6.3 в строке под номером k=3. Далее полученная строка умножается на коэффициент c_{43} = 1/5 и вычитается из строки с номером i = 4 таблицы 6.3.

Результат вычитания – единственная строка – помещается под номером i = 4 в таблицу 6.4. Наконец поделим строку с i = 4 таблицы 6.4 на коэффициент c_{44} = 5/4 и результат запишем в ту же таблицу ниже штриховой линией под номером k=4.

Таблица 6.4

i	*j*=1	*j*=2	*j*=3	*j* =4	*j*=5	*j*=6	*j*=7	*j*=8
4	0	0	0	5/4	–3/4	1/8	1/8	1
k=4	0	0	0	1	–3/5	1/10	1/10	4/5

Таким образом заканчивается прямой ход обращения матрицы. Строки, пронумерованные индексом *k*, образовали расширенную матрицу, у которой левая половина приведена к диагональному виду (см. табл. 6.5).

Таблица 6.5

k	*j*=1	*j*=2	*j*=3	*j* =4	*j*=5	*j*=6	*j*=7	*j*=8
1	1	–1/2	1/2	1	1/2	0	0	0
2	0	1	–3/5	0	–1/5	2/5	0	0
3	0	0	1	15/4	3/4	3/8	–5/8	0
4	0	0	0	1	–3/5	1/10	1/10	4/5

Таблица 6.5 содержит коэффициенты и свободные члены системы (6.6), полученные преобразованиями прямого хода метода Гаусса. Числа в столбцах j = 1, …, 4 таблицы 6.5 можно рассматривать как коэффициенты при неизвестных элементах обратной матрицы, а числа в столбцах j = 5, …, 8 представляют собой четыре столбца свободных членов. Из уравнения (4.18) следует, что числа в строке k=4 и столбцах j = 5, …, 8 составляют последнюю строку искомой обратной матрицы. Для вычисления остальных строк следует выполнить операции обратного хода метода Гаусса.

Строка k=3 в таблице 6.5 эквивалентна системе уравнений

$$x_{3j} + (15/4)\, x_{4j} = d_{3j+4}, \qquad (6.8)$$

где j = 1,…,4, d_{3j+4} – элементы строки k=3 и столбцов j = 1,…,4 таблицы 6.5.

Уравнения (6.8) аналогичны уравнению (4.19). Предпоследняя строка обратной матрицы получится решением уравнений (6.8), которое удобно выполнить формально с помощью таблицы 6.5. Для этого

достаточно из строки k=3 вычесть строку k=4, умноженную на элемент d_{34}=15/4. Таблица 6.5 преобразуется в таблицу 6.6.

Таблица 6.6

k	j=1	j=2	j=3	j =4	j=5	j=6	j=7	j=8
1	1	−1/2	1/2	1	1/2	0	0	0
2	0	1	−3/5	0	−1/5	2/5	0	0
3	0	0	1	0	3	0	−1	−3
4	0	0	0	1	−3/5	1/10	1/10	4/5

Элементы таблицы 6.6 в строке k=3 с j = 5,…,8 являются предпоследней строкой искомой обратной матрицы.

Следующий шаг обратного хода состоит в вычитании из строки k=2 таблицы 6.6 строки k=3, умноженной на элемент d_{23} = −3/5. При этом в правой половине строки k=2 сформируется вторая строка обратной матрицы $\boldsymbol{A}^{-1}$. Результаты действий записаны в таблицу 6.7.

Таблица 6.7

k	j=1	j=2	j=3	j =4	j=5	j=6	j=7	j=8
1	1	−1/2	1/2	1	1/2	0	0	0
2	0	1	0	0	8/5	2/5	−3/5	−9/5
3	0	0	1	0	3	0	−1	−3
4	0	0	0	1	−3/5	1/10	1/10	4/5

На последнем шаге приходится решать уравнения

$$x_{1j} + d_{12}\, x_{2j} + d_{13}\, x_{3j} + d_{14}\, x_{4j} = d_{1j+4} \quad (j = 1,\ldots,4), \tag{6.9}$$

где d_{kj} – элементы таблицы 6.7, а x_{kj} (k = 2, 3, 4) – уже ранее вычисленные элементы строк 2, 3, 4 обратной матрицы. Для этого из первой строки таблицы 6.7 вычитается сумма остальных строк, каждый элемент которой предварительно умножается на d_{1j}, где j – номер столбца (j=1,…,4).

В результате получается расширенная матрица, приведенная в таблице 6.8.

Таблица 6.8

k	j=1	j=2	j=3	j =4	j=5	j=6	j=7	j=8
1	1	0	0	0	2/5	1/10	1/10	−1/5
2	0	1	0	0	8/5	2/5	−3/5	−9/5
3	0	0	1	0	3	0	−1	−3
4	0	0	0	1	−3/5	1/10	1/10	4/5

Левая часть таблицы 6.8 содержит единичную матрицу, правая часть представляет собой искомую обратную матрицу $\boldsymbol{A}^{-1}$:

$$\boldsymbol{A}^{-1} = \begin{bmatrix} 2/5 & 1/10 & 1/10 & -1/5 \\ 8/5 & 2/5 & -3/5 & -9/5 \\ 3 & 0 & -1 & -3 \\ -3/5 & 1/10 & 1/10 & 4/5 \end{bmatrix} \qquad (6.10)$$

Заметим, что перед расчетом $\boldsymbol{A}^{-1}$ не вычислялся детерминант исходной матрицы det $\boldsymbol{A}$. Если бы det $\boldsymbol{A}$ = 0, то обратная матрица не существовала бы. В этом случае на одном из шагов прямого хода столбец, из которого берется очередной ведущий элемент, состоял бы только из нулей. Эта особенность указывалась в § 4.3 при описании метода Гаусса. Для независимой проверки вычислим детерминант матрицы (6.7) любым из способов, приведенных в предыдущей главе, и получим det **A** = −10. Следовательно, матрица, обратная данной, существует.

Для проверки результата можно рассчитать обратную матрицу другим способом, описанным в начале параграфа, который выражает матрицу $\boldsymbol{A}^{-1}$ формулой (6.5). Однако вычислять 16 определителей 3-го порядка плюс еще один 4-го порядка – занятие трудоемкое. Рациональнее провести перемножение полученной матрицы (6.10) и исходной (6.7). Произведение этих матриц оказывается равно единичной матрице, в чем может убедиться каждый читатель. Следовательно, матрица (6.10) обратна данной (6.7), согласно определению (6.1).

В приведенном выше примере обращение матрицы демонстрировалось с помощью восьми таблиц. Столь подробное

изложение было проведено для увеличения наглядности используемого алгоритма Гаусса.

Использование компьютеров привело к распространению более эффективных методов. Например, алгоритм Гаусса–Жордана позволяет провести обращение неособенной матрицы размером $n \times n$ за n шагов. Сначала строится расширенная матрица размером $n \times 2n$, левая половина которой является исходной матрицей $\boldsymbol{A}$, а правая половина – единичной матрицей.

На каждом k-м шаге алгоритма k-я строка расширенной матрицы делится на ведущий элемент. Элементы остальных строк (с номерами $i \neq k$) преобразуются по закону

$$c_{ij}^{(k)} = c_{ij}^{(k-1)} - c_{ik}^{(k-1)} \; c_{kj}^{(k)} \; . \qquad (6.11)$$

Пример 2. Кратко проиллюстрируем алгоритм Гаусса-Жордана на матрице (6.7), обращенной в примере 1. Приведем только преобразуемые расширенные матрицы на каждом шаге алгоритма и соответствующие ведущие элементы:

$$\boldsymbol{B_0} = \begin{bmatrix} 2 & -1 & 1 & 2 & 1 & 0 & 0 & 0 \\ 1 & 2 & -1 & 1 & 0 & 1 & 0 & 0 \\ 3 & 0 & -1 & -3 & 0 & 0 & 1 & 0 \\ 1 & -1 & 1 & 3 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$c_{11} = 2;$$

$$\boldsymbol{B_1} = \begin{bmatrix} 1 & -1/2 & 1/2 & 1 & 1/2 & 0 & 0 & 0 \\ 0 & 5/2 & -3/2 & 0 & -1/2 & 1 & 0 & 0 \\ 0 & 3/2 & -5/2 & -6 & -3/2 & 0 & 1 & 0 \\ 0 & -1/2 & 1/2 & 2 & -1/2 & 0 & 0 & 1 \end{bmatrix},$$

$$c_{22}^{(1)} = 5/2;$$

$$\boldsymbol{B_2} = \begin{bmatrix} 1 & 0 & 1/5 & 1 & 2/5 & 1/5 & 0 & 0 \\ 0 & 1 & -3/5 & 0 & -1/5 & 2/5 & 0 & 0 \\ 0 & 0 & -8/5 & -6 & -6/5 & -3/5 & 1 & 0 \\ 0 & 0 & 1/5 & 2 & -3/5 & 1/5 & 0 & 1 \end{bmatrix},$$

$$c_{33}^{(2)} = -8/5;$$

$$\boldsymbol{B_3} = \begin{bmatrix} 1 & 0 & 0 & 1/4 & 1/4 & 1/8 & 1/8 & 0 \\ 0 & 1 & 0 & 9/4 & 1/4 & 5/8 & -3/8 & 0 \\ 0 & 0 & 1 & 15/4 & 3/4 & 3/8 & -5/8 & 0 \\ 0 & 0 & 0 & 5/4 & -3/4 & 1/8 & 1/8 & 0 \end{bmatrix},$$

$$c_{44}^{(3)} = 5/4;$$

$$\boldsymbol{B_4} = \begin{bmatrix} 1 & 0 & 0 & 0 & 2/5 & 1/10 & 1/10 & -1/5 \\ 0 & 1 & 0 & 0 & 8/5 & 2/5 & -3/5 & -9/5 \\ 0 & 0 & 1 & 0 & 3 & 0 & -1 & -3 \\ 0 & 0 & 0 & 1 & -3/5 & 1/10 & 1/10 & 4/5 \end{bmatrix}.$$

Видно, что на 4-м шаге исходная расширенная матрица преобразовалась в матрицу, у которой левая половина является единичной матрицей, а правая половина – матрицей, обратной данной (6.7), т.е. совпадающей с (6.10).

Для освоения других способов обращения матриц следует ознакомиться с более подробными курсами численных методов, указанных в списке литературы [1,3,16,17].

Глава 7. РЕШЕНИЕ НЕЛИНЕЙНЫХ УРАВНЕНИЙ

§ 7.1. Выделение корней

В настоящей главе рассматриваются нелинейные уравнения, которые можно привести к стандартному виду:

$$f(x) = 0, \tag{7.1}$$

где $f(x)$ – произвольная функция действительного переменного x.

Корнем уравнения (7.1) называется число ξ, которое, будучи значением аргумента $x = \xi$, обращает уравнение (7.1) в тождество. Иначе говоря, при $x = \xi$ функция $f(x)$ принимает нулевое значение:

$$f(x = \xi) \equiv 0. \tag{7.2}$$

Следовательно, проблема поиска корней сводится к отысканию нулей функции $f(x)$. Выяснение наличия нулей функции $f(x)$ и определение их количества целесообразно проводить, максимально используя свойства функции $f(x)$ в левой части уравнения (7.1). Например, если функция $f(x)$ представляет собой полином n-й степени, то она не может иметь более n нулей.

Задача решения нелинейного уравнения вида (7.1) разбивается на две части. Первая часть заключается в нахождении интервалов, каждый из которых содержат внутри себя *только один корень*, т.е. интервалов, содержащих только один нуль функции $f(x)$. Вторая часть состоит в вычислении значения корня *с заданной погрешностью*.

Первая часть является в принципе более сложной, так как не существует универсальных и эффективных алгоритмов выделения интервалов, содержащих единственный корень, для функции $f(x)$ общего вида.

При поиске интервалов, содержащих единственный нуль функции $f(x)$, находящейся в левой части уравнения (7.1), полезно опираться на следующие теоремы математического анализа.

Теорема 1. Пусть функция $f(x)$ определена на интервале $[a, b]$ и непрерывна на нем. Если на концах этого интервала функция $f(x)$ имеет различные знаки, т.е.

$$f(a) \cdot f(b) < 0 \tag{7.3}$$

то внутри интервала [*a, b*] находится хотя бы один корень уравнения (7.1). Доказательство следует из непрерывности функции $f(x)$ и тождества (7.2).

Замечание. В найденном интервале может находиться несколько корней: как нечетное, так и четное количество (см. рис.7.1).

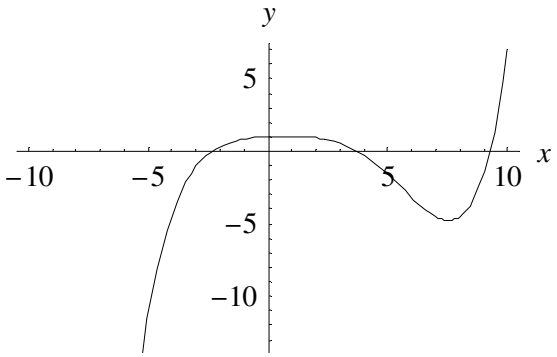


Рис. 7.1. График знакопеременной функции.
На интервале (–5, 10) данная функции имеет 3 нуля.

Теорема 2. Пусть функция $f(x)$ определена на интервале [*a, b*] и непрерывна на нем. Если на концах этого интервала производная $f'(x)$ сохраняет знак и при этом для функции $f(x)$ выполняется условие знакопеременности (7.3), то внутри интервала [*a, b*] находится единственный корень уравнения (7.1).

Доказательство от противного.

Таким образом, для нахождения интервалов, содержащих корни уравнения (7.1), целесообразно исследовать области знакопеременности функции $f(x)$ и ее производной $f'(x)$.

Пример 1. Решить уравнение $x \cdot ln(x) - 1 = 0$.

Вычислим значения функции $f(x) = x \cdot ln(x) - 1$ в точках $x = 1$ и $x = e$, где e – основание натуральных логарифмов. Получим следующие значения:

$$f(1) = -1 < 0,\ f(e) = e - 1 > 0.$$

Производная $f'(x) = ln(x) + 1$, т.е. положительна при $x \geq 1$. Следовательно, интервал $[1, e]$ содержит один корень. Нетрудно доказать, что функция $f(x) = x \cdot ln(x) - 1$ имеет единственный нуль в своей области определения.

При наличии современной компьютерной техники теоретический анализ иногда целесообразно сочетать с просмотром значений функции $f(x)$ на интервале, границы которого задаются пользователем. Отображение значений функции возможно как в текстовом, так и в графическом режиме. Если функция $f(x)$ достаточно гладкая, то такой предварительный просмотр позволит довольно быстро выявить интервалы знакопеременности функции $f(x)$, т.е. интервалы, содержащие корни уравнения (7.1). Для быстро осциллирующей функции или резко изменяющейся в некоторых диапазонах следует просматривать более узкие интервалы.

После того как найдены границы интервала, содержащего только один нуль функции $f(x)$, следует применять какой-либо численный метод нахождения корня уравнения (7.1) на этом интервале. Наиболее простые и достаточно универсальные методы изложены в следующих параграфах настоящей главы. Эти методы работают достаточно быстро и устойчиво, но требуют определенного начального приближения x_0 искомого корня. Для получения начального приближения x_0, как правило, необходимо знать границы интервала, содержащего искомый корень.

При постановке задачи решения нелинейного уравнения вида (7.1) задается допустимая погрешность $\varepsilon > 0$ вычисления корня. Следовательно, в процессе решения ищется не абсолютно точное, а приближенное значение корня. Приемлемым решением называется любое число x, которое удовлетворяет неравенству

$$|x - \xi| \leq \varepsilon,$$

где ξ – точное значения корня заданного уравнения (7.1).

К сожалению, быстродействие современных компьютеров иногда провоцирует студентов на примитивный способ простого перебора. При этом пользователь вычисляет значения функции $f(x)$ с шагом аргумента, равным величине погрешности ε. Процессор современного персонального компьютера, совершающий миллионы операций в секунду, позволит даже таким возмутительно неэффективным способом найти интервал знакопеременности шириной ε. В качестве искомого

корня берется середина этого интервала. Конечно, этот результат является приближенным значением корня с заданной точностью, но про подобные действия уже было сказано одним знаменитым человеком: «Низкий класс, нечистая работа».

§ 7.2. Метод половинного деления

Простым, устойчивым и достаточно эффективным является ***метод половинного деления***, называемый также ***бисекцией*** или ***дихотомией***.

Пусть нам дано уравнение (7.1). Используя методику, изложенную в § 7.1, находим интервал знакопеременности [*a, b*], содержащий корень данного уравнения. Алгоритм вычислительного процесса бисекции изображен на рис. 7.2 в форме блок-схемы.

Принцип работы алгоритма очевиден из приведенной схемы, поэтому можно ограничиться краткими комментариями. В самом начале задается погрешность вычисления корня ε и величина допустимой невязки δ. Далее вводятся границы интервала [*a, b*], содержащего искомый корень, и проверяется свойство знакопеременности функции $f(x)$ на концах заданного интервала. При отсутствии знакопеременности придется изменить границы начального интервала [*a, b*], в противном случае начинается итерационный процесс. На каждом шаге текущий интервал делится на две равные части точкой *c*. Эта точка является одной из границ нового, вдвое более узкого, интервала знакопеременности функции $f(x)$. Процесс продолжается до тех пор, пока интервал знакопеременности не станет у́же заранее заданной погрешности вычисления корня ε.

Кроме того, учитывается, что функция $f(x)$ в области своего нуля может иметь очень большую производную. Тогда в найденной точке $c \approx \xi$ функция $f(x)$ может иметь значение $f(c)$, значительно отличающееся от нуля. Поэтому в алгоритм добавляется условие, из-за которого уменьшение интервала, содержащего корень, продолжается до тех пор, пока абсолютная величина $|f(c)|$ не станет меньше заранее заданного малого числа δ. За приближенное значение искомого корня принимается средняя точка последнего интервала знакопеременности.

Данный алгоритм легко программируется и успешно реализуется на современных компьютерах.

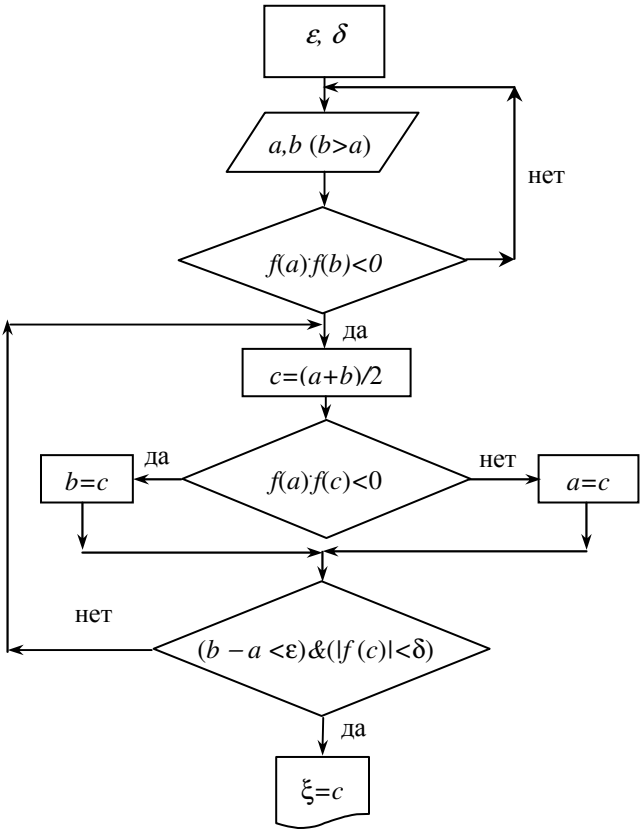


Рис. 7.2. Алгоритм схемы половинного деления.

a, b – границы интервала, в котором происходит поиск корня, ε и δ – заданные погрешности для корня и значения функции в точке найденного значения корня, ξ – значение искомого корня.

Метод половинного деления дает последовательность приближений, сходящихся к точному значению корня, если функция $f(x)$ непрерывна на предварительно выделенном интервале $[a, b]$. Следует заметить, что

монотонность функции $f(x)$ на этом интервале, вообще говоря, не является обязательной.

Более быстрые алгоритмы, т.е. требующие меньшего количества вычислений для достижения заданной точности, рассмотрены в следующих параграфах этой главы.

§ 7.3. Метод Ньютона

Пусть на интервале $[a, b]$ функция $f(x)$ непрерывна, первая и вторая производные $f'(x)$ и $f''(x)$ непрерывны и монотонны. Пусть x_n – некоторое приближенное значение корня из интервала $[a, b]$. Тогда точное значение корня можно представить в виде

$$\xi = x_n + h_n, \tag{7.4}$$

где h_n – малая поправка.

Разложим функцию $f(x)$ в точке $x = x_n$ в ряд Тейлора по степеням малой величины h_n до линейного члена:

$$f(x_n + h_n) \approx f(x_n) + h_n \cdot f'(x_n). \tag{7.5}$$

Так как $f(x_n + h_n) = f(\xi) = 0$, то величина поправки выразится из уравнения (7.5) таким образом:

$$h_n = -f(x_n) / f'(x_n). \tag{7.6}$$

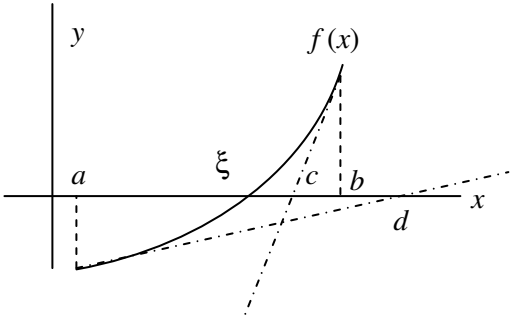
Следовательно, очередное приближение корня можно выразить в виде рекуррентного соотношения:

$$x_{n+1} = x_n - f(x_n) / f'(x_n). \tag{7.7}$$

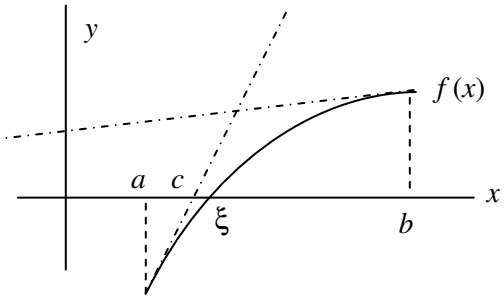
Исследования рекуррентного соотношения показали, что итерации по формуле (7.7) сходятся к истинному значению корня, если за начальное приближение x_0 взять ту границу исходного знакопеременного интервала $[a, b]$, для которой выполняется следующее условие:

$$f(x_0) \cdot f''(x_0) > 0 \tag{7.8}$$

Вместо аналитического доказательства последнего утверждения приведем графические иллюстрации (см. рис. 7.3).



а



б

Рис. 7.3. Иллюстрация сходимости процесса Ньютона:
а) $f''(x) > 0$ на всем интервале $[a, b]$,
б) $f''(x) < 0$ на всем интервале $[a, b]$.

На рис. 7.3.а изображен случай, когда вторая производная $f''(x) > 0$ на всем интервале $[a, b]$, $f(a) < 0$, $f(b) > 0$. Примем за начальное приближение корня $x_0 = a$. Величина производной $f'(x_0)$ равна тангенсу

угла наклона касательной к графику функции $f(x_0)$ в точке x_0. Следовательно, первая поправка к корню (7.6) на рис. 7.3.а изобразится отрезком ad. Точка d (следующее приближение корня x_1) лежит вне интервала $[a, b]$. Следовательно, дальнейшее использование итераций по формуле (7.7) некорректно.

Если за начальное приближение корня принять $x_0 = b$, то первый шаг процесса (7.7) даст следующее приближение корня $x_1 = c$, которое лежит внутри интервала $[a, b]$. При этом видно, что дальнейшие приближения x_2, x_3, x_4 … будут располагаться слева от истинного значения корня ξ. Последовательность приближений x_n сходится к ξ справа налево вдоль числовой оси x.

На рис. 7.3.б приведен другой случай, когда на всем интервале $[a, b]$ вторая производная $f''(x) < 0$. Аналогичные рассуждения вновь показывают справедливость условия сходимости (7.8). В данном случае последовательность приближений x_n ($n = 1, 2, \ldots$) сходится к истинному значению корня ξ слева направо.

После выбора начального приближения корня запускается процесс (7.7) и останавливается по достижении требуемой погрешности.

Ясно, что для использования метода Ньютона необходимо, чтобы производная $f'(x)$ нигде на интервале поиска корня не равнялась нулю.

Свойства рекуррентного соотношения таковы, что в общем случае условие $|x_n - x_{n-1}|<\varepsilon$ не обеспечивает выполнения неравенства $|x_n - \xi|<\varepsilon$. В курсе численных методов получено следующее соотношение для оценки погрешности достигнутого приближения искомого корня x_n:

$$| x_n - \xi | \le \frac{M_2}{2m_1} (x_n - x_{n-1})^2, \qquad (7.9)$$

где M_2 – максимум абсолютной величины второй производной $f''(x)$ на интервале $[a, b]$:

$$M_2 = \max_{a \le x \le b} \left| f''(x) \right|, \qquad (7.10)$$

m_1 – минимум абсолютной величины первой производной $f'(x)$ на том же интервале:

$$m_1 = \min_{a \le x \le b} \left| f'(x) \right|. \qquad (7.11)$$

Если, как и в предыдущем параграфе, задано максимально допустимое отклонение δ значения функции от нуля в точке приближенного корня x_n, то следует дополнительно проверять условие

$$| f(x_n) | < \delta \tag{7.12}$$

и при необходимости продолжать итерации по формуле (7.7).

В качестве искомого значения корня берется последнее приближение $x_n \approx \xi$.

Важнейшим достоинством метода Ньютона является высокая скорость сходимости. Ясно, что количество шагов алгоритма Ньютона до достижения значения $x_{n+1} \approx \xi$ существенно зависит от поведения производной $f'(x)$ на интервале $[a, b]$. Но практика показывает, что в большинстве случаев количество шагов алгоритма (7.7), по крайней мере, на порядок меньше, чем в методе половинного деления.

Недостатками метода Ньютона являются следующие: 1) необходимость вычисления первой производной функции на каждом шаге процесса, 2) требование монотонности второй производной на интервале поиска корня, 3) необходимость вычисления второй производной на концах интервала $[a, b]$.

Кроме того, применение данного метода становится «опасным» в тех случаях, когда вблизи корня первая производная близка к нулю. При этом очередная поправка к корню может быть большой и разложение (7.5) до линейного члена станет неудовлетворительным. В этих случаях очередное значение корня может даже выйти из интервала $[a, b]$, что приведет к расходимости процесса (7.7).

Примечание. Из рис. 7.3 видно, что последовательные приближения искомого корня получаются с помощью прямых линий, которые касательны к графику функция $f(x)$. По этой причине метод Ньютона часто называется методом касательных.

§ 7.4. Метод секущих

Этот метод весьма сходен с методом Ньютона, но на каждом шаге итераций вместо производной используется разделенная разность

$$[f(x_n) - f(x_{n-1})] / [x_n - x_{n-1}]. \tag{7.13}$$

Подстановка выражения (7.13) вместо производной в уравнение (7.7) дает рекуррентное соотношение для поиска корня уравнения (7.1) в следующем виде:

$$x_{n+1} = x_n - f(x_n) \cdot [x_n - x_{n-1}] / [f(x_n) - f(x_{n-1})]. \qquad (7.14)$$

В качестве двух начальных приближений корня x_0 и x_1 удобно брать границы исходного интервала $[a, b]$. Иллюстрация применения метода секущих приведена на рис.7.4.

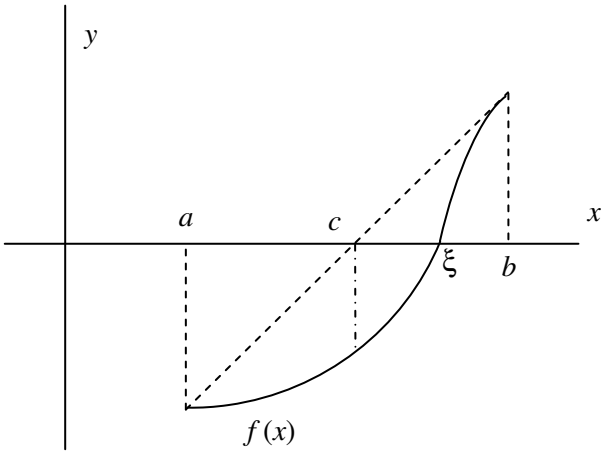


Рис. 7.4. Иллюстрация сходимости метода секущих.
ξ – точка истинного корня. В качестве начальных приближений взяты границы исходного интервала $x_0 = a$ и $x_1 = b$. Следующим приближением корня является точка $x_2 = c$.

Для сходимости процесса (7.14) требуется лишь непрерывность функции $f(x)$. Справедливости ради следует заметить, что при выделении интервала, содержащего единственный корень, ранее нам пришлось использовать монотонность первой производной $f'(x)$. Кроме того, ясно, что если на каком-либо шаге процесса (7.14) окажется $f(x_n) = f(x_{n-1})$, то произойдет сбой алгоритма.

Условия сходимости, как правило, обеспечиваются на первом этапе решения уравнения (7.1) при выборе интервала знакопеременности функции $f(x)$. При постоянстве знака второй производной $f''(x)$ на интервале поиска корня процесс (7.14) сходится почти так же быстро, как и процесс Ньютона (7.7).

Для оценки погрешности достигнутого приближения корня можно воспользоваться очевидным неравенством:

$$| x_n - \xi | \le \frac{f(x_n)}{m_1}, \qquad (7.15)$$

где величина m_1 определена формулой (7.11).

В курсе вычислительной математики доказано, что при использовании метода секущих выполнение условия $|x_n - x_{n-1}| \le \varepsilon$ гарантирует неравенство $|x_n - \xi| \le \varepsilon$, т.е. достижение заданной погрешности вычисления искомого корня ξ.

Скорость сходимости метода секущих, вообще говоря, несколько ниже, чем у метода Ньютона, что искупается гораздо более общими условиями сходимости. Это значит, что количество шагов процесса (7.14) может быть несколько больше, чем у процесса (7.7). Но при использовании рекуррентной формулы (7.14) не приходится многократно вычислять значения первой производной $f'(x)$ на каждом шаге итерации и, можно вообще не рассматривать поведение второй производной $f''(x)$ на интервале $[a, b]$.

Пример 2. Для сопоставления методов Ньютона и секущих рассмотрим простейшее нелинейное уравнение $x^2 - 2 = 0$. Легко обнаруживается интервал, содержащий один корень: $a=1$, $b=2$. Производная $f'(x)=2x$ монотонная на интервале $[0, 1]$, вторая производная $f''(x)$ является положительной константой. Процесс Ньютона следует начинать с точки $x_0=1$, в качестве начальных приближений метода секущих возьмем границы интервале $[0, 1]$. Результаты четырех первых итераций приведены в следующей таблице.

Таблица 7.1

Номер итерации n	x_n (метод Ньютона)	x_n (метод секущих)
1	1,500000000	1,333333333
2	1,416666667	1,400000000
3	1,414215683	1,414634146
4	1,414213452	1,414211438

Наблюдается быстрая сходимость обоих процессов к точному значению корня $\xi = \sqrt{2} = 1{,}414213562\ldots$ Метод Ньютона при том же числе шагов дает несколько меньшие отклонения от точного значения искомого корня ξ по сравнению с методом секущих. С другой стороны, при использовании метода Ньютона приходится на каждом шаге итерации дополнительно вычислять значения производной $f'(x_n)$.

Пример 3. Решение важной физической задачи о положении максимума спектра теплового излучения приводит к нелинейному уравнению вида:

$$5\, e^{-X} + X - 5 = 0. \tag{7.16}$$

где X – безразмерная переменная.

Корень X_0 уравнения (7.16) связан с физическими величинами следующим соотношением:

$$X_0 = \frac{2\pi\hbar c}{\beta\ k_B}, \tag{7.17}$$

где $\hbar$ – постоянная Планка, k_B – постоянная Больцмана, c – скорость света, β – константа смещения Вина. Константа β связывает абсолютную температуру T излучающего черного тела и длину волны λ_{max} электромагнитного излучения, для которой спектр теплового излучения имеет максимум:

$$\lambda_{max}\, T = \beta. \tag{7.18}$$

Решив нелинейное уравнение (7.16), следует выразить из (7.17) параметр β через найденный корень X_0 и вычислить значение размерной

константы смещения Вина. Эти операции предлагается выполнить читателям самостоятельно и сравнить результат с табличным значением, приведенным в любом добротном справочнике физических констант.

ГЛАВА 8. ЧИСЛЕННОЕ ИНТЕГРИРОВАНИЕ

§ 8.1. Принцип построения квадратурных формул

Пусть задана функция $f(x)$ на интервале $[a, b]$. Необходимо вычислить определенный интеграл:

$$I(f, a, b) = \int_a^b f(x)dx . \qquad (8.1)$$

Методы численного интегрирования используются, когда первообразная функции $f(x)$ существует, но не представляется через известные элементарные или специальные функции.

Задача численного интегрирования может формулироваться в следующих вариантах.

В первом варианте исходная (подынтегральная) функция задана только в виде набора значений y_i для конечного количества узлов x_i ($x_i \in [a, b]$), не обязательно равноотстоящих.

Во втором варианте известен алгоритм вычисления значения функции $f(x)$ для произвольного аргумента x из интервала $[a, b]$. Ясно, что второй вариант сводится к первому, но при этом можно выбирать месторасположение узлов x_i внутри интервала интегрирования. Ниже будет показано, что специальное распределение узлов внутри интервала $[a, b]$ позволяет значительно уменьшить погрешность вычисления определенного интеграла (8.1).

Кроме того, численное интегрирование применяется, когда первообразная функции $f(x)$ выражается через специальные функции очень громоздким образом. Значения спецфункций вычисляются приближенно с помощью соответствующих рядов, и поэтому иногда методы численного интегрирования дают результат с меньшей погрешностью при тех же затратах компьютерного времени.

Обычный подход к проблеме численного интегрирования заключается в замене подынтегральной функции $f(x)$ на некоторую приближенную, которую можно проинтегрировать аналитически. В качестве этой приближенной функции может быть использована интерполяционная или аппроксимирующая функция.

Чаще всего приближенная функция выбирается в виде полинома, из-за простоты его интегрирования. Первообразная полинома степени m есть полином степени $(m + 1)$.

Рассмотрим ситуацию, когда подынтегральная функция задана набором значений y_i для произвольных узлов x_i, принадлежащих интервалу $[a, b]$, причем $i = 0, 1, \ldots, n$. Оба вышеупомянутых варианта сводятся к данной постановке задачи.

Согласно изложенному в главе 2, для произвольного набора $(n+1)$ точек x_i , y_i $(i = 0, 1, \ldots, n)$ существует единственный интерполяционный полином Лагранжа (2.3). Этот полином $L_n(x)$ подставим в (8.1) на место подынтегральной функции. При этом мы получаем приближенное значение определенного интеграла:

$$I(f, a, b) \approx \int_a^b L_n(x)dx \, . \tag{8.2}$$

Применим теорему Ньютона–Лейбница и представим результат в виде

$$I(f, a, b) \approx \sum_{i=0}^{n} A_i y_i \, , \tag{8.3}$$

где A_i $(i = 0, 1, \ldots, n)$ – постоянные коэффициенты, полученные в результате интегрирования полинома Лагранжа (2.3) и подстановки пределов:

$$A_i = \int_a^b \ell_i(x)dx \, , \tag{8.4}$$

где функции ℓ_i $(i = 0, 1, \ldots, n)$ представляются произведениями (2.7). Заметим, что функции ℓ_i $(i = 0, 1, \ldots, n)$ представляют собой полиномы n-й степени, полностью определяемые положением узлов x_i $(i = 0, 1, \ldots, n)$ на интервале интегрирования.

Следовательно, для приближенного вычисления определенного интеграла требуется вычислить сумму (8.3) произведений значений функции y_i в узлах x_i на коэффициенты A_i, определяемые формулой

(8.4). Сумма, стоящая в правой части равенства (8.3), называется квадратурной, а выражение (8.3) называется ***приближенной квадратурной формулой.***

Таким образом, основная трудность численного интегрирования сводится к нахождению коэффициентов A_i для квадратурной формулы (8.3). Заметим, что значения коэффициентов A_i (i = 0, 1, …, n), вычисляемых согласно (8.4) с помощью выражений (2.7), не зависят от вида интегрируемой функции $f(x)$. Видно, что произведения (2.7) зависят только от расположения узлов x_i (i = 0, 1, …, n) внутри интервала интегрирования [a, b]. Из-за единственности интерполяционного полинома Лагранжа $L_n(x)$, проведенного через заданные точки интегрируемой функции (x_i , y_i), значение определенного интеграла, рассчитанного по квадратурной формуле (8.3), определяется только расположением узлов x_i (i = 0, 1, …, n).

Проводить практическое вычисление коэффициентов A_i квадратурной формулы (8.3) путем интегрирования полиномов n-й степени ℓ_i, которые представляются произведениями (2.7), довольно затруднительно. Гораздо более простой способ основан на том, что для подынтегральных функций $f(x) = x^k$, где k = 0, 1, …, n, квадратурная формула (8.3) является точной. Иначе говоря, равенства

$$\int_a^b x^k dx = \sum_{i=0}^{n} A_i y_i , \quad k = 0, 1, \ldots, n \tag{8.5}$$

справедливы для любого расположения узлов x_i (i = 0, 1, …, n) на интервале интегрирования [a, b].

С другой стороны, определенные интегралы в правых частях уравнений (8.5) вычисляются по формуле Ньютона–Лейбница:

$$\int_a^b x^k dx = \frac{b^{k+1} - a^{k+1}}{k+1}, \quad k = 0, 1, \ldots, n . \tag{8.6}$$

Приравнивая правые части соответствующих уравнений (8.5) и (8.6), получим n + 1 уравнений для вычисления n + 1 искомых коэффициентов

A_i (i = 0, 1, …, n). Систему полученных уравнений можно записать в виде:

$$I_0 = \sum_{i=0}^{n} A_i \ , \quad I_1 = \sum_{i=0}^{n} A_i x_i \ , \ \dots \ , I_n = \sum_{i=0}^{n} A_i x_i^n \ , \tag{8.7}$$

где числовые значения правых частей уравнений (8.6) обозначены I_k (k = 0, 1, …, n).

Система уравнений (8.7) является линейной относительно неизвестных величин коэффициентов A_i (i = 0, 1, …, n). Способы решения таких систем описаны в главе 4. Таким образом, вычисляются искомые коэффициенты A_i (i = 0, 1, …, n) и квадратурная формула (8.3) готова к применению. Следует отметить, что, хотя при получении квадратурной формулы использовался интерполяционный полином Лагранжа (2.3), описанный способ численного интегрирования не требует построения этого полинома в явном виде.

Пример 1. Пусть интервал интегрирования определяется границами a = 0 и b = 1. Известны значения подынтегральной функции в узлах x_0=1/4, x_1 = 1/2, x_2 = 3/4. Подсчет величин I_k (k = 0, 1, 2) по формулам (8.6) дает числовые значения I_0 = 1, I_1 = 1/2, I_2 = 1/3. Система (8.7) примет вид:

$A_0 + A_1 + A_2 = 1,$

$A_0 / 4 + A_1 / 2 + 3 A_2 / 4 = 1/2,$

$A_0 / 16 + A_1 / 4 + 9 A_2 / 16 = 1/3.$

Решением последней системы являются следующие значения коэффициентов A_0 = 2/3, A_1 = −1/3, A_2 = 2/3. Теперь вычисление приближенного значения определенного интеграла может проводиться по следующей квадратурной формуле:

$$\int_0^1 f(x)dx \ \approx 2\, y_0 / 3 - y_1 / 3 + 2\, y_2 / 3,$$

где $y_0 = f(x_0 = 1/4)$, $y_1 = f(x_1 = 1/2)$, $y_2 = f(x_2 = 3/4)$.

Примечание. Описанный в данном параграфе способ вычисления коэффициентов квадратурной формулы базируется на замене подынтегральной функции $f(x)$ интерполирующим полиномом. Если через заданные точки x_i , y_i ($i = 0, 1, \ldots, n$) провести другую интерполирующую функцию (не полином), то приближенное вычисление определенного интеграла (8.1) также можно проводить по квадратурной формуле (8.3). Но при этом коэффициенты A_i нельзя вычислять ни по формуле (8.4), ни решением системы (8.7). Значения коэффициентов A_i и способ их вычисления зависят от вида выбранной интерполирующей функции.

§ 8.2. Квадратурные формулы Ньютона–Котеса

В предыдущем параграфе было показано, что численные значения коэффициентов A_i ($i = 0, 1, \ldots, n$) определяются положением узлов x_i (i=0, 1, …, n) на диапазоне интегрирования $[a, b]$. Выбирая узлы x_i различными способами, мы будем получать разные наборы коэффициентов A_i.

Предположим, что мы можем вычислить $n + 1$ значений подынтегральной функции $f(x)$ в произвольных $n + 1$ точках диапазона интегрирования $[a, b]$.

Рассмотрим в первую очередь равномерное распределение узлов x_i по диапазону интегрирования $[a, b]$. Для этого разобьем диапазон $[a, b]$ на n равных интервалов, каждый длиной

$$h = (b - a) / n. \tag{8.8}$$

Узлами будут границы интервалов:

$$x_i = a + i\,h, \quad (i = 0, 1, \ldots, n). \tag{8.9}$$

Ясно, что $x_0 = a$, $x_n = b$.

Вычислим (или измерим) значения подынтегральной функции в выбранных узлах $y_i = f(x_i)$, $i = 0, 1, \ldots, n$. Теперь, как и в предыдущем параграфе, заменим подынтегральную функцию $f(x)$ интерполяционным полиномом Лагранжа (2.3). Коэффициенты квадратурной формулы A_i (i=0, 1, …, n) будем вычислять по формуле (8.4) с помощью (2.7). Тогда

выражения для коэффициентов A_i могут быть представлены в следующем виде:

$$A_i = (b - a)\, H_i, \quad i = 0, 1, \ldots, n, \tag{8.10}$$

где введены новые обозначения:

$$H_i = \frac{(-1)^{n-i}}{n\ i!\ (n-i)!} \int_0^n \frac{q(q-1)\ldots(q-n)}{q-i} dq\,, \tag{8.11}$$

$$q = (x - a) / h, \quad dq = dx / h.$$

Параметры (8.11) называются ***коэффициентами Котеса***. Нетрудно убедиться, что при любом конечном n выполняются равенства:

$$H_i = H_{n-i} \qquad \text{и} \qquad \sum_{i=0}^{n} H_i = 1.$$

Таким образом, для вычисления определенного интеграла (8.1) следует применять квадратурную формулу следующего вида:

$$\int_a^b f(x)dx \approx (b - a) \sum_{i=0}^{n} H_i y_i\,. \tag{8.12}$$

Формулы (8.12), где коэффициенты H_i определены равенствами (8.11), называются ***квадратурными формулами Ньютона–Котеса*** n-го порядка.

Для практического применения целесообразно рассмотреть некоторые частные случаи.

Пусть $n = 1$. Это значит, что диапазон интегрирования $[a, b]$ содержит только два узла: $x_0 = a$ и $x_1 = b$.

Вычислим коэффициенты Котеса для этого случая. Подставляя в (8.11) $n = 1$ и $i = 0, 1$ получим:

$$H_0 = -1 \int_0^1 \frac{q(q-1)}{q} dq = 1/2, \quad H_1 = 1 \int_0^1 q\, dq = 1/2. \qquad (8.13)$$

Подстановка коэффициентов (8.13) в (8.12) дает квадратурную формулу Ньютона–Котеса первого порядка:

$$I(f, a, b) = \int_{x_0}^{x_1} f(x)dx \approx h\,(y_0 + y_1)/2 = (b-a)\,(y_0 + y_1)/2\,. \qquad (8.14)$$

В данном случае величина шага h равна длине всего диапазона интегрирования $[a, b]$.

При таким приближении подынтегральная функция $f(x)$ заменяется линейной (полиномом Лагранжа 1-й степени). Графически значение определенного интеграла $I(f, a, b)$ изображается площадью криволинейной трапеции, ограниченной сверху (или снизу) графиком функции $f(x)$ (см. пример на рис. 8.1).

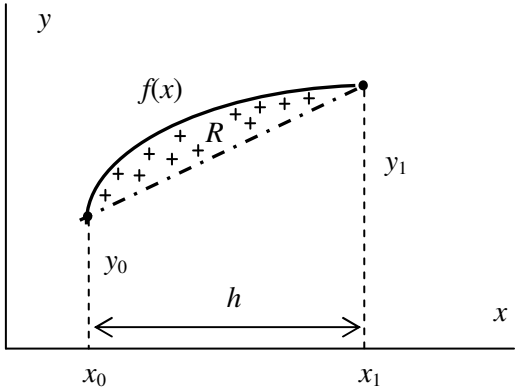


Рис. 8.1. К выводу формулы трапеций.
Сплошная линия – график интегрируемой функции $f(x)$, штрихпунктирная – график линейного приближения интегрируемой функции на интервале $[x_0, x_1]$. Крестиками заполнена область R, площадь которой равна погрешности квадратурной формулы (8.14).

Приближенное значение интеграла, выраженное квадратурной суммой (8.14), равно площади обычной трапеции с той же высотой h. Площадь криволинейного сектора дает погрешность R вычисления определенного интеграла с помощью квадратурной формулы (8.14).

Когда данный способ вычисления определенного интеграла используют на практике, то обычно диапазон интегрирования $[a, b]$ разбивают на N одинаковых поддиапазонов шириной

$$h = (b - a) / N \tag{8.15}$$

и к каждому из них применяют формулу (8.14). Тогда искомый интеграл представится суммой, каждый член которой будет содержать общий множитель (8.15). Каждое значение интегрируемой функции y_k ($k = 0, 1, \ldots, N$) в эту сумму будет входить дважды, кроме двух крайних y_0 и y_N. В результате интеграл $I(f, a, b)$ выразится следующей суммой:

$$I(f, a, b) \approx h \{ \sum_{k=0}^{N-1} y_k + \frac{y_0 + y_N}{2} \}. \tag{8.16}$$

Последняя формула называется ***формулой трапеций***.

Теперь пусть $n = 2$. Это значит, что диапазон интегрирования $[a, b]$ содержит три узла: $x_0 = a$, $x_1 = (a + b) / 2$, $x_2 = b$. В данном случае шаг интегрирования (8.8) равен половине диапазона интегрирования $h = (b - a) / 2$.

Коэффициенты Котеса в данном случае получаются вычислением интегралов (8.11) с подстановкой $n = 2$ для $i = 0, 1, 2$:

$$\begin{aligned} H_0 &= \frac{1}{2}\,\frac{1}{2}\int_0^2 (q-1)(q-2)\,dq = \frac{1}{6}, \\ H_1 &= -\frac{1}{2}\,\frac{1}{2}\int_0^2 q(q-2)\,dq = \frac{2}{3}, \\ H_2 &= \frac{1}{2}\,\frac{1}{2}\int_0^2 q(q-1)\,dq = \frac{1}{6}. \end{aligned} \tag{8.17}$$

Искомый интеграл представится в виде:

$$I(f, a, b) = \int_{x_0}^{x_2} f(x)dx \approx h\,(y_0 + y_1)/2 = (b-a)\,(\frac{1}{6}\,y_0 + \frac{2}{3}\,y_1 + \frac{1}{6}\,y_2), \quad (8.18)$$

где $(b - a) = 2h$.

Формула (8.18) была получена заменой подынтегральной функции $f(x)$ на квадратичную, проходящую через заданные точки $(x_i\ ,\ y_i)$, $i = 0, 1, 2$. Полученное значение искомого интеграла графически изображается площадью под параболой (см. рис. 8.2).

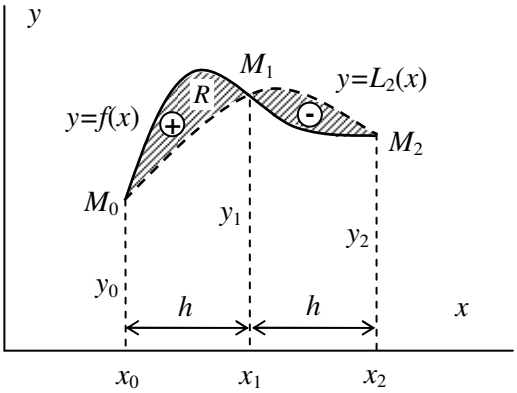


Рис. 8.2. К выводу формулы Симпсона.
Жирная сплошная линия – график интегрируемой функции $f(x)$, штриховая – график квадратичного приближения $L_2(x)$ интегрируемой функции на интервале $[x_0, x_2]$. Заштрихованная область R выражает погрешность квадратурной формулы (8.18). Видно, что на интервале $[x_0, x_1]$ квадратичное приближение дает заниженный результат, на интервале $[x_1, x_2]$ – завышенный.

На практике весь диапазон интегрирования $[a, b]$ обычно предварительно разбивается на четное число $N = 2M$ равных интервалов, которые попарно объединяются в M поддиапазонов. Каждый поддиапазон содержит три узла: два крайних и один

внутренний. К каждому из M поддиапазонов применяется формула (8.18), т.е. каждой тройке узлов соответствует своя интерполирующая парабола. Параболы стыкуются в совпадающих крайних узлах отдельных поддиапазонов. Номера всех таких узлов четные от $i = 2$ до $i = 2M - 2$, поэтому при суммировании выражений (8.18) по поддиапазонам следует учесть, что слагаемые с этими номерами встретятся дважды.

В результате суммирования с учетом значений коэффициентов Котеса (8.17) получается рабочая формула

$$I(f, a, b) \approx \frac{h}{3}(y_0 + 4\,S_O + 2\,S_E + y_{2M}), \tag{8.19}$$

где для краткости введены обозначения следующих сумм:

$$S_O = y_1 + y_3 + \ldots + y_{2M-1}, \quad S_E = y_2 + y_4 + \ldots + y_{2M-2}. \tag{8.20}$$

Заметим, что сумма S_O содержит M слагаемых, а сумма S_E состоит из $M-1$ слагаемых. Шаг интегрирования равен $h=(b-a)/(2M)=(b-a)/N$.

Формула (8.19) называется ***квадратурной формулой Симпсона***.

При $n = 3$ аналогичным путем получается квадратурная формула с тремя узлами на интервале интегрирования:

$$I(f, a, b) = \int_{x_0}^{x_3} f(x)dx \approx \frac{3}{8}h\,(y_0 + 3y_1 + 3y_2 + y_3), \tag{8.21}$$

где $h = (b - a) / 3$.

При решении прикладных задач диапазон интегрирования $[a, b]$ предварительно разбивается на равные поддиапазоны, число которых кратно трем ($N = 3M$), и к каждому из поддиапазонов применяется формула (8.20). Тогда получится рабочая формула вида:

$$I(f, a, b) \approx \frac{3}{8}h\,(y_0 + 3\,S_1 + 3\,S_2 + 2\,S_3 + y_N). \tag{8.22}$$

Шаг интегрирования равен $h = (b - a) / (3M) = (b - a) / N$, а отдельные суммы выражены следующими формулами:

$$S_1 = \sum_{k=0}^{M-1} y_{3k+1}\,, \quad S_2 = \sum_{k=0}^{M-1} y_{3k+2}\,, \quad S_3 = \sum_{k=1}^{M-1} y_{3k}\,. \tag{8.23}$$

Сумма (8.22) называется ***квадратурной формулой Ньютона*** или ***формулой «трех восьмых»***. В практике квадратурная формула Ньютона употребляется гораздо реже, чем формулы Симпсона и трапеций.

Еще реже применяется ***квадратурная формула Боде***, которая получается из общей формулы (8.12) подстановкой $n = 4$. Интервал интегрирования содержит 5 равноотстоящих узлов $x_i = a + i\,h$, $i = 0, 1, \ldots, 4$. Приближенное значение определенного интеграла выразится следующей суммой:

$$I(f, a, b) \approx \frac{2}{45}\,h\,(7y_0 + 32y_1 + 12y_2 + 32y_3 + 7y_4), \tag{8.24}$$

где $h = (b - a) / 4$.

Читателю предоставляется возможность самостоятельно записать рабочую формулу численного интегрирования при разбиении диапазона интегрирования $[a, b]$ на N равных поддиапазонов.

§ 8.3. Квадратурная формула Гаусса

Все формулы численного интегрирования дают результат с некоторой погрешностью. Оценки погрешностей приводятся в следующем параграфе. Здесь заметим, что уменьшение погрешности расчета является одной из важнейших задач численных методов.

В предыдущем параграфе рассматривались различные квадратурные формулы для равномерного расположения узлов в интервале интегрирования $[a, b]$. Оказывается, при фиксированном количестве узлов можно добиться значительного уменьшения погрешности вычисления определенного интеграла, если отказаться от их равномерного расположения.

Для нахождения оптимального расположения узлов (которое обеспечивает уменьшение погрешности численного интегрирования) прежде всего сведем интервал интегрирования общего вида $[a, b]$ к

стандартному интервалу [–1, 1] с помощью линейной замены переменной интегрирования:

$$x = (a + b) / 2 + t\,(b - a) / 2. \tag{8.25}$$

Исходный интеграл (8.1) приобретет вид:

$$\int_a^b f(x)dx \;=\; \int_{-1}^{1} f\left(\frac{a+b}{2} + \frac{b-a}{2}t\right)dt\,, \tag{8.26}$$

где t – новая переменная интегрирования.

В соответствии с общей квадратурной формулой (8.3) запишем:

$$\int_{-1}^{1} f(t)\;dt \;\approx\; \sum_{i=1}^{n} A_i f(t_i)\,. \tag{8.27}$$

Нашей целью является выбор таких n узлов t_i внутри интервала интегрирования [–1, 1] и таких n коэффициентов A_i (i = 1, …, n), чтобы формула (8.27) была *точной* для всех *полиномов* максимально большой степени. Мы имеем возможность варьировать $2n$ величин t_i и A_i (i = 1, …, n), следовательно, эта максимальная степень полинома равна $N = 2n - 1$.

Далее требуется доказать следующую лемму.

Лемма. Для достижения поставленной цели необходимо и достаточно, чтобы формула (8.27) выполнялась точно для следующих $2n$ подынтегральных функций $f(t) = 1, t, t^2, \ldots, t^{2n-1}$.

Доказательство.

По условиям леммы выполняются следующие равенства для степенных функций $f(t) = t^k$ (k = 0, 1, …, $2n - 1$):

$$\int_{-1}^{1} t^k dt \;=\; \sum_{i=1}^{n} A_i t_i^k\,, \quad k = 0, 1, \ldots, 2n - 1. \tag{8.28}$$

Как было принято выше, подынтегральную функцию $f(t)$ представим в виде полинома степени $2n - 1$:

$$f(t) = \sum_{k=0}^{2n-1} C_k t^k \,, \tag{8.29}$$

где C_k – коэффициенты полинома.

Проведем интегрирование полинома (8.29) по интервалу [–1, 1], используя условие (8.28) и изменяя порядок суммирования:

$$\int_{-1}^{1} f(t)dt = \sum_{k=0}^{2n-1} C_k \int_{-1}^{1} t^k dt = \sum_{k=0}^{2n-1} C_k \sum_{i=1}^{n} A_i t_i^k =$$

$$= \sum_{i=1}^{n} A_i \sum_{k=0}^{2n-1} C_k t_i^k = \sum_{i=1}^{n} A_i f(t_i) \,. \tag{8.30}$$

Последнее равенство обусловлено определением (8.29).

Мы получили, что равенство (8.27) является точным для полинома вида (8.29), при этом в ходе преобразований мы воспользовались условиями (8.28).

Лемма доказана.

Эта лемма обеспечивает справедливость следующего важного утверждения. Пусть нам удастся найти такие числа t_i и A_i (i = 1, …, n), обращающие формулу (8.27) в точную, используя в качестве подынтегральных степенные функции вида t^k (k = 0, 1, …, 2n – 1). Тогда эти найденные числа t_i и A_i обеспечат точность формулы (8.27) и для любой подынтегральной функции, которая представляется в виде полинома степени 2n – 1.

Теперь в уравнениях (8.28) проинтегрируем аналитически левые части:

$$\int_{-1}^{1} t^k dt = \begin{cases} \dfrac{2}{k+1}, (k - \textit{четно}) \\ 0, (k - \textit{нечетно}) \end{cases} . \tag{8.31}$$

Сравнение (8.31) с правыми частями уравнений (8.28) дает нам систему 2n уравнений для 2n искомых неизвестных t_i , A_i (i = 1, …, n):

$$\sum_{i=1}^{n} A_i = 2, \quad \sum_{i=1}^{n} A_i\, t_i = 0,$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \qquad (8.32)$$
$$\sum_{i=1}^{n} A_i\, t_i^{2n-2} = \frac{2}{2n-1}, \quad \sum_{i=1}^{n} A_i\, t_i^{2n-1} = 0.$$

Система является нелинейной относительно неизвестных t_i , A_i (i = 1, …, n). Для ее решения целесообразно воспользоваться свойствами полиномов Лежандра. Краткие сведения о полиномах Лежандра приведены в приложении 2.

Выберем подынтегральные функции $f(x)$ в виде:

$$f(x) = t^k P_n(t),\ k = 0, 1, \ldots, n-1, \qquad (8.33)$$

где $P_n(t)$ - полином Лежандра степени n.

Это значит, что все n функций (8.33) являются полиномами степени $\le 2n-1$. Для таких подынтегральных функций, согласно доказанному выше, формула (8.27) является точной и коэффициенты A_i (i = 1, …, n) удовлетворяют системе (8.32).

Применим квадратурную формулу (8.27) для функций (8.33):

$$\int_{-1}^{1} t^k P_n(t)\ dt = \sum_{i=1}^{n} A_i\, t_i^k P_n(t_i), \quad k = 0, 1, \ldots, n-1. \qquad (8.34)$$

В силу свойства ортогональности полиномов Лежандра левые части всех уравнений (8.34) равны нулю. Следовательно, и правые части (8.34) будут равны нулю

$$\sum_{i=1}^{n} A_i\, t_i^k P_n(t_i) = 0 \qquad (8.35)$$

при произвольных значениях A_i, если положить $P_n(t_i) = 0$.

Таким образом, оказывается, что в качестве узлов t_i следует взять точки нулей полиномов Лежандра степени n. Точки нулей полиномов

Лежандра различных степеней рассчитаны с большой точностью и сведены в специальные таблицы. Характерным свойством узлов t_i и коэффициентов A_i в данном методе является симметрия их значений относительно центра таблицы $t = 0$.

Если числа t_i известны, то система (8.32) становится линейной относительно величин A_i. Для ее решения можно применить любой метод, описанный в главе 4. Численные значения узлов t_i и коэффициентов A_i с точностью до восьми знаков для различных n приведены в таблице приложения 3.

Вычислив значения узлов t_i и коэффициентов A_i для выбранного n, можно вернуться к исходной переменной интегрирования x и записать ***квадратурную формулу Гаусса*** в следующем виде:

$$\int_a^b f(x)dx \approx \frac{b-a}{2} \sum_{i=1}^{n} A_i y_i , \qquad (8.36)$$

где

$$y_i = f(x_i) \text{ и } \quad x_i = (a + b) / 2 + t_i (b - a) / 2. \qquad (8.37)$$

Точность интегрирования методом Гаусса резко возрастает с увеличением числа узлов в интервале интегрирования [a, b]. Теоретические оценки и расчеты показывают, что для интервалов небольшой величины обычно достаточная точность обеспечивается при $n = 8$. Соответствующая таблица узлов t_i и коэффициентов A_i приведена в приложении 3.

Если диапазон интегрирования [a, b] имеет значительную ширину, его можно разбить на несколько равных поддиапазонов и провести процедуру вычисления интеграла методом Гаусса для каждого поддиапазона отдельно. Интеграл по j-му поддиапазону с границами [a_j , b_j] с помощью линейного преобразования (8.37) представляется в виде суммы, аналогичной (8.36):

$$I (f, a_j , b_j) \approx \frac{b_j - a_j}{2} \sum_{i=1}^{n} A_i y_i , \qquad (8.38)$$

где $y_i = f(x_i)$, $x_i = (a_j + b_j) / 2 + t_i (b_j - a_j) / 2$.

Искомое значение определенного интеграла по полному диапазону [a, b] вычисляется как сумме интегралов (8.38) по всем поддиапазонам.

§ 8.4. Погрешности квадратурных формул

Погрешность любой формулы численного интегрирования определяется остаточным членом, равным разности точного значения определенного интеграла и квадратурной суммы вида (8.3):

$$R = \int_a^b f(x)dx - \sum_{i=0}^{n} A_i y_i \ . \tag{8.39}$$

Так как точное значение интеграла неизвестно, то остаточный член приходится вычислять приближенно с помощью известных значений подынтегральной функции $y_i = f(x_i)$.

За погрешность численного интегрирования с помощью квадратурной формулы можно принять приближенное абсолютное значение остаточного члена (8.39) или его оценку сверху.

Обратимся вначале к методу трапеций. Интервал интегрирования [*a, b*] разбивается на *N* равных поддиапазонов шириной *h*, и к каждому из них применяется формула (8.14). Остаточный член (8.39) для любого поддиапазона может быть представлен в виде:

$$R = -\frac{h^3}{12} f''(\xi), \tag{8.40}$$

где ξ – некоторая точка внутри данного поддиапазона. Рисунок 8.1 показывает, что в случае использования метода трапеций абсолютная величина остаточного члена (8.39) равна площади криволинейного сектора. По рис. 8.1 хорошо видно, что при $f''(x) < 0$ на интервале интегрирования (x_0, x_1) квадратурная формула (8.14) дает значение с недостатком и остаточный член (8.40) имеет положительное значение. При $f''(x) > 0$ формула (8.14) дает значение определенного интеграла с избытком и величина остаточного члена *R* согласно (8.40) отрицательна.

Интеграл по всему диапазону [*a, b*] дается квадратурной формулой трапеций (8.16). Погрешность этой формулы определяется суммой остаточных членов (8.40) по всем *N* поддиапазонам. Оценка сверху абсолютной величины остаточного члена формулы (8.16) может быть записана в следующем виде:

$$R = h^2 \frac{b-a}{12} M_2, \tag{8.41}$$

где M_2 – максимум абсолютного значения производной 2-го порядка функции $f(x)$ на интервале интегрирования $[a, b]$:

$$M_2 = \max_{a \le x \le b} \left| f''(x) \right|. \tag{8.42}$$

Если отсутствует аналитическое выражение второй производной $f''(x)$, то для оценки величины M_2 используется представление конечной разностью (1.3) или формула численного дифференцирования, приводимая в следующей главе.

Перейдем к оценке погрешности численного интегрирования по методу Симпсона. В этом методе диапазон интегрирования $[a, b]$ разбит на четное число $N = 2M$ равных интервалов шириной h, к каждой паре из которых применяется квадратурной формула (8.18). Остаточный член этой формулы представляется так:

$$R = -\frac{h^5}{90} f^{(\mathrm{IV})}(\xi), \tag{8.43}$$

где ξ – некоторая точка внутри данного поддиапазона.

Численное интегрирование методом Симпсона по всему диапазону $[a, b]$ дается квадратурной формулой (8.19). Погрешность этой формулы определяется суммой остаточных членов (8.43) по всем $M = N/2$ поддиапазонам. Приближенное значение погрешности можно записать в виде:

$$R = h^4 \frac{b-a}{180} M_4, \tag{8.44}$$

где M_4 – максимум абсолютного значения производной 4-го порядка $f^{(\mathrm{IV})}(x)$ подынтегральной функции $f(x)$ на интервале $[a, b]$:

$$M_4 = \max_{a \le x \le b} \left| f^{(IV)}(x) \right|. \tag{8.45}$$

Численная оценка величины $f^{(\mathrm{IV})}(x)$ при отсутствии аналитического вида функции $f(x)$ весьма затруднительна. Поэтому на практике оценку погрешности вычисления определенного интеграла методом Симпсона реализуют эмпирическим путем. Интеграл (8.1) вычисляют по формуле (8.19) дважды: с малым шагом h_1 и с вдвое меньшим шагом $h_2 = h_1 / 2$. Соответствующие значения квадратурных сумм (8.19) обозначим I_1 и I_2. Тогда в качестве оценки искомого интеграла берется величина I_2.

Если интегрируемая функция достаточно гладкая, то ее четвертая производная $f^{(\mathrm{IV})}(x)$ на диапазоне $[a, b]$ изменяется медленно и ее можно приближенно заменить константой. Тогда оценка погрешности метода Симпсона представима в виде:

$$R \approx |I_2 - I_1| / 15. \tag{8.46}$$

Расчет погрешностей для квадратурных формул Ньютона–Котеса более высоких порядков приводится в специальных курсах численных методов, так как эти формулы применяются гораздо реже, чем формулы трапеций и Симпсона.

Погрешность значения определенного интеграла по квадратурной формуле Гаусса также определяется величиной остаточного члена этой формулы. Остаточный член квадратурной формулы Гаусса резко зависит от количества узлов n в отдельном интервале интегрирования $[a, b]$. В общем виде остаточный член, определяющий погрешность, представляется следующим выражением:

$$R_n = \frac{(b-a)^{2n+1}(n!)^4}{(2n!)^3(2n+1)} f^{(2n)}(\xi), \tag{8.47}$$

где $\xi \in [a, b]$. Например, для $n = 3$ и для $n = 6$ получаем:

$$R_3 = \frac{1}{15750}\left(\frac{b-a}{2}\right)^7 f^{(6)}(\xi) \text{ и } R_6 = \frac{1}{648984486150}\left(\frac{b-a}{2}\right)^{13} f^{(12)}(\xi).$$

Однако необходимость численных оценок производных высоких порядков делает формулу (8.47) мало пригодной для практики. Оценки погрешности приходится выполнять эмпирически, постепенно удваивая количество поддиапазонов интегрирования.

Пример 2. Для сравнения точности различных квадратурных формул рассмотрим в качестве примера вычисление значения функции *erf*(*x*=1). Эта функция задана интегралом, где аргумент является верхним пределом:

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2)\, dt. \qquad (8.48)$$

Вычисление $erf(x = 1)$ должно проводиться численно, так как для первообразной функции $exp(-t^2)$ не существует аналитического выражения.

Расчет значения определенного интеграла $\frac{2}{\sqrt{\pi}} \int_0^1 \exp(-t^2)\, dt$ проведем тремя способами: по 2-точечной формуле трапеций (8.14), по 3-точечной формуле Симпсона (8.18) и по 2-точечной формуле Гаусса (8.36). При этом диапазон интегрирования не будем разбивать на поддиапазоны. Получим:

$$I_T = \frac{1}{\sqrt{\pi}} (1 + e^{-1}) = 0{,}7717433\ldots ,$$

$$I_S = \frac{1}{3\sqrt{\pi}} (1 + 4\, e^{-1/4} + e^{-1}) = 0{,}8431028\ldots ,$$

$$I_G = \frac{1}{\sqrt{\pi}} \left\{ \exp\left[-\frac{1}{4}\left(1 - \frac{1}{\sqrt{3}}\right)^2 \right] + \exp\left[-\frac{1}{4}\left(1 + \frac{1}{\sqrt{3}}\right)^2 \right] \right\} = 0{,}8424419\ldots .$$

где I_T, I_S и I_G — значения определенного интеграла, вычисленные методами трапеций, Симпсона и Гаусса соответственно.

Точный расчет значения функции $erf(x = 1)$ с точностью до 10^{-8} проводится с помощью специального ряда и дает 0,84270079… .

Сравнение результатов показывает, что значение, вычисленное по квадратурной формуле трапеций, значительно отличается от точного результата. Квадратурная формула Гаусса дает меньшую погрешность по сравнению с формулой Симпсона.

Конечно, уменьшить погрешность любого метода можно, разбивая диапазон [*a, b*] интегрирования на поддиапазоны и используя формулы

(8.16), (8.19) и (8.38). Однако различие в точности методов остается тем же.

Вернемся к задаче вычисления функции $erf(x = 1)$. В таблице 8.1 представлены результаты численного интегрирования методом трапеций (8.16) по диапазону [0, 1] для разного числа интервалов.

Таблица 8.1

Число интервалов интегрирования	Значение квадратурной суммы
2	0,82526295
4	0,83836777
8	0,84161922
16	0,84243051
32	0,84263323
64	0,84268390
128	0,84269657

Видно, что при увеличении количества поддиапазонов метод трапеций демонстрирует весьма медленную сходимость к точному значению определенного интеграла. Поэтому метод трапеций обычно используют для быстрых приближенных оценок, которые содержат малое количество правильных значащих цифр.

Точность квадратурной формулы Гаусса выше по сравнению с формулой Симпсона при заданном разбиении диапазона интегрирования. Это позволяет, применяя метод Гаусса, при заданной погрешности вычисления определенных интегралов разбивать диапазон интегрирования на меньшее число подинтервалов, что приводит к значительному сокращению времени расчета. Опыт показывает, что количество поддиапазонов при использовании квадратурной формулы Гаусса может быть взято, грубо говоря, на порядок меньше, чем при применении метода Симпсона.

Примечание. В данном пособии не рассматриваются методы вычисления многомерных и несобственных интегралов. По этим вопросам авторы отсылают читателей к более подробным курсам численных методов, например к [3, 9].

Глава 9. ЧИСЛЕННОЕ ДИФФЕРЕНЦИРОВАНИЕ

§ 9.1. Дифференцирование интерполяционных полиномов

Вычисление производных заданных функций проводится обычно по хорошо известным правилам и формулам дифференцирования. При этом получается аналитическое выражение для производной функции. Подстановка в полученное выражение любого аргумента, принадлежащего области допустимых значений, позволяет вычислить искомое значение производной функции.

Численные (неаналитические) методы вычисления значений производной функции привлекаются в следующих случаях, встречающихся на практике.

Во-первых, когда исходная функция задана лишь таблицей значений в узлах. Промежуточные значения данной функции (для аргументов, не совпадающих с узлами) вычисляются методами интерполяции или аппроксимации.

Во-вторых, когда исходная функция определяется некоторым алгоритмом, который позволяет получить значение $y = f(x)$ для любого аргумента x, принадлежащего области определения функции, но аналитическое выражение (формула) для функции $f(x)$ отсутствует. Следовательно, стандартные формулы дифференцирования неприменимы.

В-третьих, зависимость $y = f(x)$ может выражаться сложной формулой с использованием специальных функций. Нахождение аналитического представления производной сопряжено с изрядными трудностями, а формула $f'(x)$ приобретает весьма громоздкий вид, мало удобный для вычислений. В частности, выражение может представляться бесконечным рядом неэлементарных функций, значения которых в принципе вычисляются приближенно. Во многих подобных случаях численные методы дифференцирования имеют преимущество – они позволяют получить значение производной с достаточной точностью за меньшее время по сравнению с расчетом по полученной формуле $f'(x)$.

Прежде чем переходить к конкретным методам, следует обратить внимание на принципиальную трудность численного дифференцирования. В курсе математического анализа доказано, что для сколь угодно близких непрерывных функций расстояние между их производными может быть неограниченно велико. Это означает, что

задача дифференцирования на пространстве непрерывных функций в принципе некорректна. Вышеуказанное свойство функций обусловливает ограниченную точность всех формул численного дифференцирования.

Целью методов численного дифференцирования является расчет значения производной $f'(x)$ функции $y = f(x)$ в определенной точке x.

Обычные методы численного дифференцирования базируются на имеющемся алгоритме вычисления значений исходной функции $y = f(x)$ для заданного аргумента x. В первом из перечисленных случаев функция $f(x)$ заменяется интерполяционной или аппроксимирующей.

Проще всего приближенное выражение производной получить с помощью соответствующей конечной разности. Пользуясь определением производной функции, запишем:

$$f'(x) = dy/dx \approx \frac{f(x+h) - f(x)}{h}, \tag{9.1}$$

где h – конечная величина. Числитель представляет собой конечную разность первого порядка (1.1), следовательно,

$$f'(x) \approx \Delta^1 y / h. \tag{9.2}$$

Таким образом, для получения приближенного значения производной в точке x необходимо вычислить два значения исходной функции $f(x)$ – в точках x и $x + h$. Аналогично можно приближенное значение производной $f'(x)$ в точке x выразить с помощью значений исходной функции $f(x)$ в точках $x – h$ и x:

$$f'(x) \approx \frac{f(x) - f(x-h)}{h}. \tag{9.3}$$

Представление производной $f'(x)$ формулами (9.1) или (9.3) означает, что на интервале $[x, x + h]$ или $[x – h, x]$ исходная функция $f(x)$ заменяется линейной.

Более точные численные оценки производной функции можно получить, используя замену исходной функции $f(x)$ интерполяционными полиномами.

Рассмотрим ситуацию, когда в нашем распоряжении имеется конечное число значений исследуемой функции $f(x)$ в равноотстоящих узлах:

$$x_i = a + i\,h, \quad y_i = f(x_i), \quad i = 0, 1, \ldots, n. \tag{9.4}$$

Теперь заменим исходную функцию $f(x)$ первым полиномом Ньютона (2.21). Для дальнейших преобразований полезно раскрыть скобки в числителях всех слагаемых:

$$P_n(x) = y_0 + q\Delta^1 y_0 + \frac{q^2 - q}{2}\cdot\Delta^2 y_0 + \frac{q^3 - 3q^2 + 2q}{6}\Delta^3 y_0 +$$
$$+ \frac{q^4 - 6q^3 + 11q^2 - 6q}{24}\Delta^4 y_0 + \frac{q^5 - 10q^4 + 35q^3 - 50q^2 + 24q}{120}\Delta^5 y_0 + \ldots, \tag{9.5}$$

где $q = (x - x_0) / h$.

Производную функции $y = f(x)$ по аргументу x представим как производную сложной функции:

$$\frac{dy}{dx} = \frac{dy}{dq}\,\frac{dq}{dx} = \frac{1}{h}\,\frac{dy}{dq}. \tag{9.6}$$

Дифференцирование полинома (9.5) по переменной x, с учетом (9.6), дает формулу для приближенного значения производной:

$$f'(x) \approx [\Delta^1 y_0 + \frac{2q-1}{2}\Delta^2 y_0 + \frac{3q^2 - 6q + 2}{6}\Delta^3 y_0 +$$
$$\frac{2q^3 - 9q^2 + 11q - 3}{12}\Delta^4 y_0 +$$
$$+ \frac{5q^4 - 40q^3 + 105q^2 - 100q + 24}{120}\Delta^5 y_0 + \ldots] / h. \tag{9.7}$$

Если продифференцировать по переменной x полином (9.7), то мы получим формулу для приближенного значения второй производной:

$$f''(x) \approx [\ \Delta^2 y_0 + (q-1)\Delta^3 y_0 + \frac{6q^2 - 18q + 11}{12}\Delta^4 y_0 +$$

$$+ \frac{2q^3 - 12q^2 + 21q - 10}{12}\Delta^5 y_0 + \dots\ ] / h^2. \qquad (9.8)$$

Напомним, что все конечные разности в формулах (9.2), (9.7) и (9.8) должны вычисляться для аргумента $x = x_0$. При практическом использовании формул (9.5) и (9.6) в качестве узла x_0 выбирается узел таблицы, ближайший слева к значению аргумента x.

Если требуется вычислить значение первой или второй производной в каком-либо узле таблицы (9.2), то для этого достаточно в выражениях (9.5) и (9.6) положить $q = 0$. Тогда получим следующие полезные формулы численного дифференцирования:

$$f'(x_0) \approx [\ \Delta^1 y_0 - \Delta^2 y_0 / 2 + \Delta^3 y_0 / 3 - \Delta^4 y_0 / 4 + \Delta^5 y_0 / 5\ \dots] / h. \qquad (9.9)$$

$$f''(x_0) \approx [\ \Delta^2 y_0 - \Delta^3 y_0 + \frac{11}{12}\Delta^4 y_0 - \frac{5}{6}\ \Delta^5 y_0 + \dots] / h^2. \qquad (9.10)$$

В выражениях (9.5) и (9.7) – (9.10) приведены несколько первых членов полинома Ньютона и его производных. В зависимости от числа узлов таблицы (9.5) и требуемой точности количество слагаемых в (9.7) – (9.10) может быть уменьшено или увеличено. Заметим, что конечные разности $\Delta^i y_0$ в формулах (9.9) и (9.10) должны вычисляться для точки x_0, в которой ищутся производные.

Видно, что оценка производной (9.2) является частным случаем формулы (9.7), если степень полинома Ньютона положить равной единице (т.е. заменить неизвестную функцию $f(x)$ линейной).

С помощью второго полинома Ньютона (2.26) возможно построить формулы численного дифференцирования, аналогичные вышеприведенным (9.7) – (9.10). Также в качестве дифференцируемых функций можно брать интерполяционные полиномы Стирлинга (2.31) и Бесселя (2.32).

Погрешности формул (9.5) – (9.8) определяются абсолютной величиной первого из отбрасываемых членов. Так как конечная разность n-го порядка $\Delta^n y \sim h^n$, то ясно, что погрешности оценок производных (9.1) и (9.3) имеют порядок $\sim h$, т.е. являются весьма грубыми приближениями.

Пример 1. Некоторая функция $f(x)$ задана таблицей значений 9.1.

Таблица 9.1

i		0,1	0,2	0,3	0,4	0,5	0,6	0,7
i		0,4 992	0,9 933	1,4 776	1,9 471	2,3 971	2,8 232	3,2 211

Пусть требуется вычислить первую и вторую производную данной функции $f(x)$ в точке $x = 0{,}24$. Видно, что функция $f(x)$ задана таблицей с равноотстоящими узлами, шаг таблицы $h = 0{,}1$. Узел 0,2 является ближайшим слева к заданному значению аргумента $x = 0{,}24$ и поэтому выбирается за x_0. Для вычисления производных используем формулы (9.7) и (9.8), построенные на основе первого интерполяционного полинома Ньютона.

Предварительно рассчитаем необходимые конечные разности пяти порядков в точке $x = x_0$ и сведем их в таблицу 9.2.

Таблица 9.2

	x_i	y_i	Δy_i	$\Delta^2 y_i$	$\Delta^3 y_i$	$\Delta^4 y_i$	$\Delta^5 y_i$
	0,2	0,9933	0,4843	-0,0148	-0,0047	0,00019	$4{,}5\cdot 10^{-5}$
	0,3	1,4776	0,4695	-0,0195	-0,0045	0,00024	
	0,4	1,9471	0,4500	-0,0240	-0,0043		
	0,5	2,3971	0,4261	-0,0282			
	0,6	2,8232	0,3979				
	0,7	3,2211					

Для заданного аргумента $x = 0{,}24$ величина $q = (x - x_0) / h = 0{,}4$. Расчет по формуле (9.7) с использованием только двух слагаемых дает величину 4,85726. Погрешность полученного значения определяется

абсолютной величиной третьего слагаемого формулы (9.7), которая равна $6{,}25 \cdot 10^{-4}$. Следовательно, производная $f'(x = 0{,}24)$ равна 4,857 с тремя верными знаками после запятой.

Для получения более точного значения следует в формуле (9.7) использовать большее количество слагаемых. Например, расчет четырех слагаемых дает значение $f'(x = 0{,}24) = 4{,}8567$ с четырьмя верными знаками после запятой. Учет следующих слагаемых в (9.7) не даст достоверных результатов, хотя бы потому, что исходные данные содержат ограниченное количество верных значащих цифр.

Расчет значения второй производной в точке $x = 0{,}24$ проводится по формуле (9.8) с использованием конечных разностей той же таблицы 9.2. Вычисление четырех слагаемых дает значение $f''(x = 0{,}24) = -1{,}188$ с тремя верными знаками после запятой.

Формулы, полученные с помощью интерполяционных полиномов Ньютона, обладают существенным недостатком. Значение производной в определенной точке x_0 вычисляется по формулам (9.7) – (9.10) с использованием значений исходной функции $f(x)$ только для аргументов $x \geq x_0$. Иначе говоря, оценка численного значения производной рассчитывается на основе информации об исходной функции $f(x)$ в области аргументов, превышающих значение x_0.

Аналогично, формулы, полученные из второго интерполяционного полинома Ньютона, также страдают односторонностью оценки, так как используют значения исходной функции $f(x)$ в точках, расположенных левее заданного аргумента. Большую точность дают более «симметричные» формулы численного дифференцирования, которые используют значения заданной функции $f(x)$ в точках $x > x_0$ и в $x < x_0$, где x_0 – аргумент, для которого требуется отыскать значение производной. Такие формулы рассматриваются в следующем параграфе.

§ 9.2. Использование разложения в ряд Тейлора

Для практических целей наиболее целесообразным является представление производных формулами, использующими значения дифференцируемой функции в точках, расположенных симметрично относительно точки, в которой должна быть вычислена производная. Такие формулы характеризуются высокой точностью, что позволяет

использовать сравнительно небольшое количество узлов для вычисления значений исходной функции $f(x)$.

Получить «симметричные» формулы численного дифференцирования можно, используя полиномы Стирлинга или Лагранжа для набора равноотстоящих узлов. Более удобным является использование разложения данной функции $f(x)$ в ряд Тейлора.

Пусть требуется найти значение производной заданной функции $f(x)$ в точке $x = x_0$. Будем полагать, что в нашем распоряжении имеется алгоритм, позволяющий получать значения функции $f(x)$ в некотором конечном интервале аргументов x в окрестности точки x_0.

Вычислим значения данной функции $f(x)$ в равноотстоящих точках

$$x_i = x_0 + i\,h, \qquad i = 0, \pm1, \pm2, \pm3, \ldots . \tag{9.11}$$

Величину h выберем малой. Тогда точки $x_1 = x_0 + h$ и $x_{-1} = x_0 - h$ можно полагать находящимися в окрестности точки x_0. Запишем следующие разложения в ряд Тейлора для функции $f(x)$:

$$f(x_1) = f(x_0) + h\,f'(x_0) + \frac{h^2}{2}\,f''(x_0) + \frac{h^3}{3!}\,f'''(x_0) + \ldots, \tag{9.12}$$

$$f(x_{-1}) = f(x_0) - h\,f'(x_0) + \frac{h^2}{2}\,f''(x_0) - \frac{h^3}{3!}\,f'''(x_0) + \ldots . \tag{9.13}$$

Сначала в разложениях (9.12) и (9.13) отбросим все члены 3-го порядка и выше. Из-за малости величины h такое приближение вполне допустимо. После этого вычтем выражение (9.13) из (9.12) и получим следующее уравнение:

$$f(x_1) - f(x_{-1}) = 2\,h\,f'(x_0).$$

Отсюда сразу получается формула для приближенного вычисления производной в точке $x = x_0$:

$$f'(x_0) \approx \frac{f(x_0 + h) - f(x_0 - h)}{2h}. \tag{9.14}$$

Погрешность последней формулы определяется величиной первого отброшенного члена разложений (9.12) и (9.13) и оценивается величиной $h^2 f'''(x_0) / 6$, т.е. имеет порядок ~ h^2. Для сравнения отметим, что погрешность формулы (9.1) на порядок больше.

Формула (9.14) называется 3-точечной оценкой первой производной функции $f(x)$, хотя для получения величины $f'(x_0)$ достаточно вычислить значения исходной функции $f(x)$ только в двух точках, симметричных относительно аргумента x_0.

Оценка производной (9.14) может быть уточнена. Для этого в разложениях (9.12) и (9.13) отбросим все члены, начиная с 4-го. После этого проведем вычитание (9.13) из (9.12) и перегруппируем члены. В результате получится уравнение, связывающее значения функции $f(x)$ и ее производных в разных точках:

$$f(x_1) - f(x_{-1}) = 2\,h\,f'(x_0) + h^3 f'''(x_0) / 3. \qquad (9.15)$$

Теперь запишем разложения в ряд Тейлора для функции $f(x)$ в точках $x_2 = x_0 + 2\,h$ и $x_{-2} = x_0 - 2\,h$, ограничиваясь членами 3-го порядка включительно:

$$f(x_2) = f(x_0) + 2\,h\,f'(x_0) + 4\frac{h^2}{2} f''(x_0) + 8\frac{h^3}{3!} f'''(x_0) + \ldots, \qquad (9.16)$$

$$f(x_{-2}) = f(x_0) - 2\,h\,f'(x_0) + 4\frac{h^2}{2} f''(x_0) - 8\frac{h^3}{3!} f'''(x_0) + \ldots. \qquad (9.17)$$

Проведем вычитание разложения (9.17) из (9.16):

$$f(x_2) - f(x_{-2}) = 4\,h\,f'(x_0) + 8\,h^3 f'''(x_0) / 3.$$

Из последнего уравнения выразим $h^3 f'''(x_0)$ и подставим в (9.15). При этом получится уравнение, которое можно разрешить относительно величины $f'(x_0)$. Таким образом, получается еще одна приближенная формула для вычисления первой производной функции $f(x)$ в точке $x=x_0$:

$$f'(x_0) \approx \frac{f(x_0-2h)-8f(x_0-h)+8f(x_0+h)-f(x_0+2h)}{12h}. \qquad (9.18)$$

Погрешность формулы (9.18) оценивается величиной $h^5 f^{(V)}(x_0) / 120$. Следовательно, погрешность формулы (9.18) на 3 порядка меньше погрешности оценки (9.14).

Формула (9.18) называется 5-точечной оценкой первой производной функции $f(x)$. Заметим, что вычисления значения функции $f(x)$ в точке $x=x_0$ не требуется.

Перейдем к выводу формул численного дифференцирования для вторых производных данной функции. Для этого вернемся к разложениям (9.12) и (9.13), отбросим все члены выше 3-го порядка и сложим их. Получим уравнение:

$$f(x_1) + f(x_{-1}) = 2f(x_0) + h^2 f''(x_0),$$

из которого непосредственно получается формула приближенного вычисления второй производной в точке $x = x_0$:

$$f''(x_0) \approx \frac{f(x_0+h)-2f(x_0)+f(x_0-h)}{h^2}. \qquad (9.19)$$

Эта формула справедливо называется 3-точечной оценкой второй производной, так как величина $f''(x_0)$ выражается через значения исходной функции $f(x)$ в трех точках, симметричных относительно узла x_0.

Погрешность оценки (9.19) также определяется отбрасываемыми членами разложения и приблизительно равна по абсолютной величине $h^2 |f^{(IV)}(x_0)| / 12$, т.е. имеет порядок ~ h^2.

Для нахождения более точной оценки второй производной необходимо воспользоваться разложениями в ряд Тейлора исходной функции до 4-го члена включительно в точках $x_1 = x_0 + h$, $x_{-1} = x_0 - h$, $x_2 = x_0 + 2h$ и $x_{-2} = x_0 - 2h$.

$$f(x_1) = f(x_0) + h f'(x_0) + \frac{h^2}{2} f''(x_0) + \frac{h^3}{3!} f'''(x_0) + \frac{h^4}{4!} f^{(IV)}(x_0), \qquad (9.20)$$

$$f(x_{-1}) = f(x_0) - h f'(x_0) + \frac{h^2}{2} f''(x_0) - \frac{h^3}{3!} f'''(x_0) + \frac{h^4}{4!} f^{(IV)}(x_0), \qquad (9.21)$$

$$f(x_2) = f(x_0) + 2h f'(x_0) + \frac{4h^2}{2} f''(x_0) + \frac{8h^3}{3!} f'''(x_0) + \frac{16h^4}{4!} f^{(IV)}(x_0), \qquad (9.22)$$

$$f(x_{-2}) = f(x_0) - 2h f'(x_0) + \frac{4h^2}{2} f''(x_0) - \frac{8h^3}{3!} f'''(x_0) + \frac{16h^4}{4!} f^{(IV)}(x_0). \qquad (9.23)$$

Сложим попарно разложения (9.20) и (9.21), а также (9.22) и (9.23). Получим уравнения:

$$f(x_1) + f(x_{-1}) = 2 f(x_0) + h^2 f''(x_0) + \frac{h^4}{12} f^{(IV)}(x_0), \qquad (9.24)$$

$$f(x_2) + f(x_{-2}) = 2 f(x_0) + 4 h^2 f''(x_0) + \frac{4h^4}{3} f^{(IV)}(x_0). \qquad (9.25)$$

Выразим произведение $h^4 f^{(IV)}(x_0)$ из (9.24) и подставим (9.25). Таким путем получим соотношение, связывающее значения функции $f(x)$ в пяти разных узлах и второй производной в точке $x = x_0$:

$$f(x_2) + f(x_{-2}) = 16\,[f(x_1) + f(x_{-1})] - 30 f(x_0) - 12 h^2 f''(x_0).$$

Отсюда получается новая формула для приближенного вычисления второй производной заданной функции в точке $x = x_0$:

$$f''(x_0) \approx$$

$$\approx \frac{-f(x_0 - 2h) + 16 f(x_0 - h) - 30 f(x_0) + 16 f(x_0 + h) - f(x_0 + 2h)}{12h^2}. \qquad (9.26)$$

Формула (9.26) называется 5-точечной оценкой второй производной, так как использует пять значений исходной функции $f(x)$ в пяти равноотстоящих точках. Численное дифференцирование по формуле дает погрешность (9.26) приблизительно равную $h^4 \mid f^{(VI)}(x_0) \mid / 90$, что на два порядка меньше погрешности оценки (9.19).

Пример 2. Вновь рассмотрим функцию, заданную таблицей 9.1 в примере 1 этого параграфа. Вычислим первую и вторую производные в узле таблицы $x = 0{,}3$, приняв это значение за центральный узел x_0. Используя величину шага $h = 0{,}1$, получим по формуле (9.14) значение первой производной $f'(x = 0{,}3) = 4{,}769$, а по формуле (9.18) – значение $f'(x = 0{,}3) = 4{,}777$. Расчеты второй производной по формулам (9.19) и (9.26) дадут значения $-1{,}48$ и $-1{,}4825$ соответственно.

Для сравнения вычислим обе производные в той же точке $x = 0{,}3$ с помощью формул (9.9) и (9.10), полученных из первого интерполяционного полинома Ньютона. Напомним, что при этом должны использоваться конечные разности, рассчитанные в точке $x = 0{,}3$, т.е. расположенные во второй строке данных таблицы 9.2. Использование двух слагаемых в сумме (9.9) дает результат $f'(x = 0{,}3) = 4{,}792$, а четырех слагаемых – значение $f'(x = 0{,}3) = 4{,}777$. Расчеты второй производной по формуле (9.10) с использованием двух и четырех слагаемых дают значения $-1{,}5$ и $-1{,}474$ соответственно.

Теперь можно открыть маленький секрет. В таблице 9.1 расположены значения функции $y = 5sin(x)$. Аналитические расчеты первой и второй производных для аргумента $x = 0{,}3$ дают значения $f'(x) = 4{,}7767$ и $f''(x) = -1{,}4776$. Сравнение результатов, полученных разными методами, позволяет сделать следующие выводы. Расчет величины $f'(x)$ по 5- точечной формуле (9.18) и с помощью выражения (9.9) с четырьмя слагаемыми дают 3 верные цифры после запятой, что для многих практических задач вполне достаточно. Расчеты по 3-точечной формуле (9.14) и по формуле (9.9) с двумя слагаемыми дают менее точные результаты.

Вычисления второй производной $f''(x)$ по 5-точечной формуле (9.26) и по формуле (9.10) с использованием четырех слагаемых дают результат с точностью до сотых долей.

Таким образом, оба подхода к вычислению производных, изложенные в § 9.1, 9.2, позволяют достичь определенной точности, однако расчеты по формулам § 9.2 значительно короче. С другой стороны, формулы § 9.1 необходимы для случаев, когда требуется вычислять производные в точках, не совпадающих с узлами.

Обращает на себя внимание более низкая точность вычислений второй производной (по сравнению с первой).

Естественным способом понижения погрешности численных расчетов величин производных кажется уменьшение шага h таблицы узлов (в тех случаях, когда это возможно). Однако принципиальная некорректность процедуры численного дифференцирования, о которой упоминалось в начале главы, приводит к иному результату. Проверка показывает [9], что с уменьшением шага h таблицы узлов точность численного дифференцирования сначала возрастает, затем начинает существенно убывать.

Обратим внимание на то, что в знаменателях всех формул численного дифференцирования стоит малое число, а это приводит к погрешностям округления при компьютерных вычислениях. При заданном шаге h относительная погрешность вычисления для второй производной больше, чем для первой.

На практике при численном расчете производных следует ограничиваться минимальным количеством значащих цифр.

§ 9.3. Численное дифференцирование при произвольном расположении узлов

Иногда требуется получить оценку значения производной k-го порядка, когда вся информация о дифференцируемой функции сводится к таблице ее значений y_i в неравноотстоящих узлах x_i ($i = 0, 1, \ldots, n$). В этих случаях по заданной таблице можно построить полином Лагранжа, а затем его аналитически дифференцировать. Однако это сопряжено с достаточно громоздкими алгебраическими преобразованиями. Альтернативой в такой ситуации является численное дифференцирование с использованием ***метода неопределенных коэффициентов***.

Пусть необходимо вычислить первую производную функции $f(x)$, заданной таблицей, в некоторой точке $x = x_U$, причем $x_0 < x_U < x_n$.

Представим значение искомой производной в интересующей нас точке x_U следующей суммой:

$$f'(x_U) = \sum_{i=0}^{n} c_i y_i + R, \qquad (9.27)$$

где величина R называется остаточным членом, а коэффициенты c_i (i= 0, 1, ... , n) пока являются неопределенными.

Для нахождения коэффициентов c_i (i = 0, 1, ... , n) положим, что остаточный член точно равен нулю, если дифференцируемая функция f (x) является степенной с показателем степени j = 0, 1, …, n:

$$f(x) = 1, x, x^2, \ldots, x^n. \tag{9.28}$$

Вычислим первые производные всех функций (9.28) в точке $x = x_U$ и подставим в равенство (9.27), полагая R = 0. Суммы в правых частях выражений (9.27) запишем, используя явный вид функций (9.28). В результате получим следующую систему из (n + 1) уравнений:

$$\sum_{i=0}^{n} c_i = 0, \quad \sum_{i=0}^{n} c_i x_i = 1, \quad \sum_{i=0}^{n} c_i x_i^2 = 2\, x_U,$$

$$\sum_{i=0}^{n} c_i x_i^3 = 3\, x_U^2, \ldots, \sum_{i=0}^{n} c_i x_i^n = n\, x_U^{n-1}. \tag{9.29}$$

Система уравнений (9.29) является линейной относительно неизвестных (n + 1) коэффициентов c_i (i = 0, 1, ... , n). Решив систему (9.29) одним из методов, описанных в главе 4, получим искомое приближенное значение первой производной в точке $x = x_U$.

Аналогично можно с помощью неопределенных коэффициентов построить алгоритм вычисления второй производной при тех же исходных данных x_i, y_i (i = 0, 1, ... , n). Для этого значение искомой второй производной $f''(x_U)$ сначала тоже представляется суммой (9.27). Численные значения коэффициентов c_i (i = 0, 1, ... , n) находятся приравниванием нулю остаточного члена R для функций (9.28). Для вычисления искомых коэффициентов получается система из (n + 1) линейных уравнений:

$$\sum_{i=0}^{n} c_i = 0, \quad \sum_{i=0}^{n} c_i x_i = 0, \quad \sum_{i=0}^{n} c_i x_i^2 = 2, \quad \sum_{i=0}^{n} c_i x_i^3 = 6\, x_U,$$

$$\sum_{i=0}^{n} c_i x_i^4 = 12\, x_U^2 \ , \ \dots \ , \ \sum_{i=0}^{n} c_i x_i^n = n\,(n-1)\ x_U^{n-2} . \qquad (9.30)$$

После получения коэффициентов c_i ($i = 0, 1, \dots , n$) из системы (9.30) приближенное значение второй производной вычисляется в виде суммы:

$$f''(x_U) \approx \sum_{i=0}^{n} c_i y_i \ .$$

Вышеизложенный метод можно распространить на вычисление производной k-го порядка. В этом случае линейная система уравнений для расчета коэффициентов c_i ($i = 0, 1, \dots , n$) записывается в следующем общем виде:

$$\sum_{i=0}^{n} c_i = 0, \quad \sum_{i=0}^{n} c_i x_i = 0, \quad \dots \ , \quad \sum_{i=0}^{n} c_i x_i^{k-1} = 0,$$

$$\sum_{i=0}^{n} c_i x_i^k = k!, \quad \sum_{i=0}^{n} c_i x_i^{k+1} = (k+1)!\, x_U, \quad \dots \ , \qquad (9.31)$$

$$\sum_{i=0}^{n} c_i x_i^n = n\,(n-1)\ \dots\ (n-k+1)\ x_U^{n-k} .$$

Пример 3. Функция задана значениями в пяти неравноотстоящих уздах, которые сведены в таблицу 9.3.

Таблица 9.3

i	0	1	2	3	4
x_i	0,1	0,3	0,4	0,7	1,2
y_i	3,664	5,466	6,677	12,17	33,07

Вычислим первую производную в точке x_U = 0,5. Для нахождения коэффициентов c_i (i = 0, 1, ... , 4) суммы (9.27) вычислим матрицу коэффициентов системы (9.29)

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0{,}1 & 0{,}3 & 0{,}4 & 0{,}7 & 1{,}2 \\ 0{,}01 & 0{,}09 & 0{,}16 & 0{,}49 & 1{,}44 \\ 0{,}001 & 0{,}027 & 0{,}064 & 0{,}343 & 1{,}728 \\ 0{,}0001 & 0{,}0081 & 0{,}0256 & 0{,}2401 & 2{,}0736 \end{bmatrix} \tag{9.32}$$

и столбец свободных членов

$$\boldsymbol{b} = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0{,}75 \\ 0{,}5 \end{pmatrix}. \tag{9.33}$$

Решение системы линейных уравнений $\boldsymbol{A}\ \boldsymbol{c} = \boldsymbol{b}$ методом Гаусса (см. главу 4) дает набор искомых коэффициентов: c_0 = 0,606; c_1 = –4,722; c_2 = 1,667; c_3 = 2,5; c_4 = –0,0505. Первая производная в заданной точке вычисляется с помощью суммы (9.27):

$$f'(x = 0{,}5) \approx \sum_{i=0}^{n} c_i y_i \ = 16{,}3. \tag{9.34}$$

Теперь вычислим вторую производную той же функции, заданной таблицей 9.3, тоже в точке x_U = 0,5. Сравнивая системы уравнений (9.29) и (9.30), замечаем, что они имеют одинаковые матрицы коэффициентов при неизвестных. Легко убедиться, что столбец свободных членов для системы (9.30) имеет вид:

$$\boldsymbol{b_2} = \begin{pmatrix} 0 \\ 0 \\ 2 \\ 3 \\ 3 \end{pmatrix}. \qquad (9.35)$$

Решением системы линейных уравнений (9.30) получаем коэффициенты для расчета второй производной: $c_0 = -5{,}556$; $c_1 = 75$; $c_2 = -88{,}89$; $c_3 = 19{,}44$; $c_4 = 0$. Затем вычисляем искомое значение: $f''(x = 0{,}5) \approx 32{,}7$.

Данный пример служит не только для демонстрации численного дифференцирования методом неопределенных коэффициентов, но и позволяет оценить точность этого метода. В таблице 9.3 были записаны значения функции $f'(x) = 3exp(2x)$. Аналитическое вычисление первой и второй производной в точке $x = 0{,}5$ показывает, что численный расчет $f'(x = 0{,}5)$ дает 3 верные значащие цифры, а погрешность величины $f''(x = 0{,}5)$ не превышает 0,003. Такую точность следует признать весьма удовлетворительной, так как таблица исходных данных содержала только 5 узлов, а значения функции y_i были заданы лишь с четырьмя значащими цифрами.

ПОСЛЕСЛОВИЕ

Для практической реализации численных методов к настоящему времени разработано обширное программное обеспечение. Широко применяются такие стандартные пакеты, как Microsoft Excel, Microcal Origin, Mathematica, MathCad, Maple, Matlab и т.д. Почти ежегодно на потребительском рынке появляются обновленные версии этих пакетов. Все современные программные средства содержат встроенную систему помощи. Может быть, следует просто открыть какой-либо пакет, реализующий численные методы, и, иногда пользуясь закладкой Help, выбрать метод и проделать надлежащие вычисления? По данному вопросу у авторов есть следующие замечания.

1. Как правило, стандартный программный пакет реализует те алгоритмы, которые сочли нужным применить сами разработчики. Далеко не все пакеты дают пользователю возможность выбирать конкретный численный метод и оценить достоверность и точность полученного результата.

2. Зачастую система помощи построена весьма лаконично. Опыт показывает, что эффективно использовать встроенную систему помощи могут лишь опытные пользователи, уже достаточно овладевшие теорией численных методов. Кроме того, как уже упоминалось в главе 2, насущной является проблема различий в российской и зарубежной терминологии численных методов.

Таким образом, совершенно не отрицая полезности существующих программных средств, авторы глубоко убеждены, что пользователь, применяя тот или иной встроенный в пакет метод, должен четко представлять себе его суть и иметь возможность контролировать точность реализации выбранного метода для каждой конкретной задачи. В ином случае пользователю следует самому реализовывать численные методы программным путем. Такие грамотные действия возможно предпринимать лишь вооружившись теоретическими знаниями по численным методам. Предлагаемое учебное пособие должно помочь читателям в решении этой задачи.

Приложение 1

КВАДРАТИЧНЫЕ ФОРМЫ

Квадратичной формой n аргументов x_i (i = 1, 2, ..., n) называется однородный полином 2-й степени следующего вида

$$U(x_1, x_2, \dots, x_n) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} x_i x_j , \qquad \text{(I.1)}$$

где a_{ij} – элементы некоторой квадратной матрицы $\boldsymbol{A}$.

Эта матрица называется *соответствующей* квадратичной форме (I.1). Матрица $\boldsymbol{A}$ является симметричной, т.е. совпадает со своей транспонированной $\boldsymbol{A}'$, т.е. $a_{ij} = a_{ji}$.

Квадратичная форма (I.1) называется *положительно определенной*, если она принимает положительные значения при любых значениях аргументов x_i (i = 1, 2, ..., n) за исключением точки $x_1 = x_2 = \dots = x_n = 0$, где величина (I.1) равна нулю.

Аналогично, квадратичная форма (I.1) называется *отрицательно определенной*, если она принимает отрицательные значения при любых значениях аргументов x_i (i = 1, 2, ... n) за исключением точки $x_1 = x_2 = \dots = x_n = 0$.

Приложение 2

ПОЛИНОМЫ ЛЕЖАНДРА

Полиномы Лежандра являются специальными функциями, которые применяются при решении многих теоретических и прикладных задач. Полином Лежандра n-й степени можно определить с помощью производной n-го порядка следующим образом:

$$P_n(z) = \frac{1}{2^n n!} \frac{d^n}{dx^n} [(z^2 - 1)^n] , \qquad n = 0, 1, 2,\dots , \quad \text{(II.1)}$$

где z – комплексная переменная.

В данном учебном пособии рассматриваются и используются полиномы Лежандра для действительного аргумента x, лежащего в интервале $x \in [-1, 1]$.

С помощью определения (II.1) легко получить явные выражения полиномов Лежандра действительного аргумента низших степеней:

$$P_0(x) = 1, \qquad P_1(x) = x, \qquad P_2(\mathrm{x}) = (3x^2 - 1) / 2,$$
$$P_3(x) = (5x^3 - 3x) / 2, \quad P_4(x) = (35x^4 - 30x^2 + 3) / 8, \dots \qquad \text{(II.2)}$$

Графики перечисленных полиномов приведены на рис.1.

Все полиномы Лежандра $P_n(x)$ имеют следующие граничные значения:

$$P_n(1) = 1, \qquad P_n(-1) = (-1)^n \qquad \text{(II.3)}$$

Нетрудно убедиться, что полиномы Лежандра четной степени являются четными функциями и наоборот.

Важным для практических применений является свойство ортогональности полиномов Лежандра:

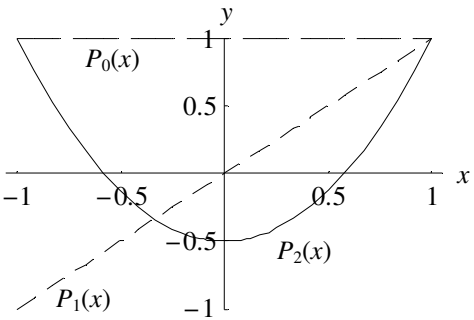
$$\int_{-1}^{1} P_n(x) \; Q_k(x) \; dx \; = 0, \qquad \text{(II.4)}$$

где $Q_k(x)$ – любой полином степени k, меньшей n ($k < n$).

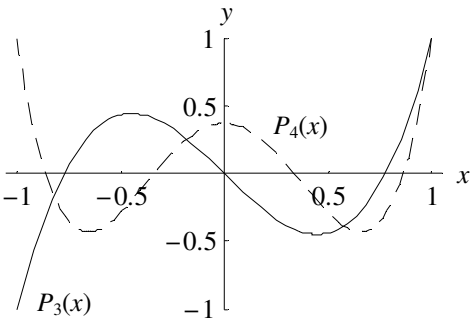
Полиномы Лежандра подчиняются рекуррентному соотношению

$$(n+1)\,P_{n+1}(x) = (2n+1)\,x\,P_n(x) - n\,P_{n-1}(x), \qquad \text{(II.5)}$$

которое, в частности, удобно для последовательного вычисления полиномы высоких степеней.



а



б

Рис.1. Графики полиномов Лежандра а) n = 0, 1, 2, б) n = 3, 4.

Приложение 3

ПАРАМЕТРЫ КВАДРАТУРНЫХ ФОРМУЛ ГАУССА

n	i	Узлы t_i	Коэффициенты A_i
2	1 ; 2	$\mp$ 0,57735027	1
3	1 ; 3	$\mp$ 0,77459667	0,55555556
	2	0	0,88888889
4	1 ; 4	$\mp$ 0,86113631	0,34785484
	2 ; 3	$\mp$ 0,33998104	0,65214516
5	1 ; 5	$\mp$ 0,90617985	0,23692688
	2 ; 4	$\mp$ 0,53846931	0,47862868
	3	0	0,56888889
6	1 ; 6	$\mp$ 0,93246951	0,17132450
	2 ; 5	$\mp$ 0,66120939	0,36076158
	3 ; 4	$\mp$ 0,23861919	0,46791394
7	1 ; 7	$\mp$ 0,94910791	0,12948496
	2 ; 6	$\mp$ 0,74153119	0,27970540
	3 ; 5	$\mp$ 0,40584515	0,38183006
	4	0	0,41795918
8	1 ; 8	$\mp$ 0,96028986	0,10122854
	2 ; 7	$\mp$ 0,79666648	0,22238104
	3 ; 6	$\mp$ 0,52553242	0,31370664
	4 ; 5	$\mp$ 0,18343464	0,36268378

Приложение 4

КРАТКИЕ ЗАМЕЧАНИЯ ОБ ИСТОЧНИКАХ ПОГРЕШНОСТЕЙ

В главах данного пособия неоднократно обсуждалась методика оценок погрешностей, возникающих вследствие использования рассматриваемых численных методов. Стремление к максимальному увеличению точности метода является естественным движением души любого начинающего исследователя. Однако прежде чем предпринимать незаурядные усилия по уменьшению погрешностей вычислений, необходимо вспомнить, что применение численного метода является лишь частью решаемой физической задачи. Поэтому вначале следует рассмотреть различные причины, влияющие на окончательный результат проводимого исследования.

Погрешности результата, полученного в ходе решения задачи, можно подразделить по причинам их возникновения. Эта классификация не является догмой, но полезна для уяснения общей ситуации с оценкой точности и надежности результата.

1. Погрешности в исходных данных задачи.
2. Погрешность, вызванная приближениями выбранной математической модели.
3. Погрешность используемого численного метода.
4. Погрешность, которая накапливается в ходе компьютерных вычислений.

Рассмотренные в данной книге погрешности относятся к третьему типу, но на точность конечного результата существенно влияют и другие причины.

Каждая из погрешностей уменьшает точность результата, но стремиться неограниченно уменьшать все погрешности нерационально и даже невозможно. Дело в том, что некоторые из погрешностей по своей природе являются «неустранимыми». Это значит, что в условиях решения данной реальной задачи не имеется возможностей их уменьшить. Например, такими часто являются погрешности первого типа.

Другие погрешности можно уменьшить, но нет смысла делать их много меньше неустранимых, особенно если это требует значительных временных или энергетических затрат.

Источниками погрешностей первого типа часто являются физические измерения. Погрешности этого типа вызываются не только

недостатком информации об исходных данных, но и использованием округленных значений констант, например таких иррациональных чисел, как π или e (основание натуральных логарифмов).

Погрешности второго типа неизбежны потому, что любое описание физического процесса является приближенным. Например, полагая силу сопротивления аэродинамической (пропорциональной квадрату скорости) или стоксовой (пропорциональной скорости), мы делаем приближение и закладываем погрешность в результат.

Аналитические методы являются в принципе точными. Численные методы для получения точного результата часто требуют бесконечного количества итераций или суммирования бесконечного числа слагаемых. На практике, разумеется, всегда используется конечное количество шагов численного метода, и тем самым обеспечивается погрешность, которая в нашей классификации отнесена к третьему типу.

Любой компьютер обрабатывает коды конечной длины. Количество разрядов машинного слова обусловливает погрешность округления. Относительная погрешность округления составляет $0{,}5 \cdot 10^{-t}$, где t – число разрядов. Эта величина кажется маленькой, но при миллионах математических операций, выполняемых процессором, окончательная погрешность вычислений (погрешность четвертого типа) может достигнуть значительной величины. Часто возникают следующие ситуации. Точность численного метода возрастает с ростом числа шагов, но на практике при прогоне тестовых задач наблюдается увеличение ошибки результата после некоторого количества итераций. Эта погрешность может возрастать, например, при машинном округлении разности близких величин или при делении на очень маленькое число. О вычислительных погрешностях имеется обширная специальная литература. Из прилагаемого списка можно рекомендовать [5, 9].

Перед получением окончательного результата необходимо оценить, хотя бы по порядку величины, все погрешности, которые могут возникнуть при решении поставленной задачи. Нет смысла тратить компьютерное время, уменьшая погрешность метода, если велика неустранимая погрешность исходных данных. В литературе описан случай, когда один специалист написал программу, реализующую сложный численный метод, который должен был обеспечить высокую точность, но в исходных данных заложил число π равным 3,14.

ЛИТЕРАТУРА

1. Демидович Б.П., Марон И.А. Основы вычислительной математики. М.: ФМ, 1963.
2. Демидович Б.П., Марон И.А., Шувалова Э.З. Численные методы анализа. М.: ФМ, 1964.
3. Березин И.С., Жидков Н.П. Методы вычислений. М.: Физматгиз, 1966.
4. Бахвалов Н.С. Численные методы. М.: Наука, 1973.
5. Каханер Д. и др. Численные методы и программное обеспечение. М.: Мир, 1998.
6. Форсайт Дж. и др. Машинные методы математических вычислений. М.: Мир, 1980.
7. Мак-Кракен Д., Дорн У. Численные методы и программирование на Фортране. М.: Мир, 1977.
8. Шуп Т. Решение инженерных задач на ЭВМ. М.: ФМ, 1982.
9. Бахвалов Н.С., Жидков Н.П., Кобельков Г.М. Численные методы. М.: Бином, 2003.
10. Данилина Н.И. и др. Численные методы. М.: ВШ, 1976.
11. Воробьева Г.Н., Данилова А.Н. Практикум по численным методам. М.: ВШ, 1979.
12. Самарский А.А., Гулин А.В. Численные методы. М.: Наука, 1989.
13. Рябенький В.С. Введение в вычислительную математику. М.: ФМ, 2000.
14. Протасов И.Д. Лекции по вычислительной математике. М.: Гелиос АРВ, 2004.
15. Кунин С. Вычислительная физика. М.: Мир, 1992.
16. Стренг Г. Линейная алгебра и ее применения. М.: Мир, 1980.
17. Хорн Р., Джонсон Ч. Матричный анализ. М.: Мир, 1989.

ОГЛАВЛЕНИЕ