МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

**САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ
ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ**

Э.В. Денисова

А.В.Кучер

Краткий курс вычислительной математики

Учебное пособие

Санкт-Петербург

2013

Э.В. Денисова, А.В. Кучер, Краткий курс вычислительной математики: Учебно-методическое пособие. – СПб: СПбГУ ИТМО, 2013. – 90с.

В пособии изложено содержание семестрового учебного курса «Вычислительная математика». Учебный курс читается в 1 семестре второго года обучения для слушателей кафедры информатики и прикладной математики и кафедры вычислительной техники. Для освоения данного учебного курса необходимо знание языка программирования высокого уровня и владение практическими навыками программирования.

Учебно-методическое пособие адресовано студентам обучающимся по специальностям:

231000.62 - Программная инженерия
231000.62.01 - Разработка программно-информационных систем
230100.62 - Информатика и вычислительная техника
230100.62.01 - Вычислительные машины, комплексы, системы и сети

Рекомендовано к печати Советом факультета Компьютерных Технологий и Управления, протокол № 3 от 12.03.2013г.

В 2009 году Университет стал победителем многоэтапного конкурса, в результате которого определены 12 ведущих университетов России, которым присвоена категория «Национальный исследовательский университет». Министерством образования и науки Российской Федерации была утверждена программа его развития на 2009–2018 годы. В 2011 году Университет получил наименование «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики»

СОДЕРЖАНИЕ

Глава 1 ПРАВИЛА ПРИБЛИЖЕННЫХ ВЫЧИСЛЕНИЙ И ОЦЕНКА ПОГРЕШНОСТЕЙ ПРИ ВЫЧИСЛЕНИЯХ 4
§ 1. Приближенные числа, их абсолютные и относительные погрешности .. 4
§ 2. Устойчивость. Корректность. Сходимость 6
§ 3. Сложение и вычитание приближенных чисел 7
§ 4. Умножение и деление приближенных чисел 8
§ 5. Погрешности вычисления значений функции 10
§ 6. Определение допустимой погрешности аргументов по допустимой погрешности функции 12
Глава 2 ВЫЧИСЛЕНИЕ ЗНАЧЕНИЙ ФУНКЦИИ 15
§ 1. Вычисление значений многочлена. Схема Горнера 15
§ 2. Вычисление значений некоторых трансцендентных функций с помощью степенных рядов 16
§ 3. Некоторые многочленные приближения 22
§ 4. Применение цепных дробей для вычисления значений трансцендентных функций 24
§ 5. Применение метода итераций для приближённого вычисления значений функций 26
Глава 3 РЕШЕНИЕ НЕЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ .. 33
§ 1. Уравнения с одним неизвестным. Метод деления пополам. Метод хорд. Метод касательной. Метод простой итерации. 33
§2. Действительные и комплексные корни алгебраических уравнений. 40
§3 Системы уравнений. Метод простой итерацию. Метод Ньютона. 41
Глава 4 РЕШЕНИЕ СИСТЕМ ЛИНЕЙНЫХ УРАВНЕНИЙ 45
§1. Прямые методы. Метод Гаусса. Метод главных диагоналей. Определитель и обратная матрица. Метод прогонки. 49
§2. Итерационные методы. Уточнение решения. Метод простой итерации. Метод Гаусса-Зейделя 56
§3. Задачи на собственные значения. Метод вращений. Трехдиагональные матрицы. 62
Глава 5 ПРИБЛИЖЕНИЕ ФУНКЦИЙ 69
§1.Точечная аппроксимация. Равномерное приближение. 69
§2. Многочлены Чебышева. Вычисление многочленов. Рациональные приближения. 70
§3. Интерполирование. Линейная и квадратичная интерполяция. Многочлен Лагранжа. Многочлен Ньютона. Кубические сплайны. Точность интерполяции. 76
§4.Аппроксимация. Метод наименьших квадратов. Эмпирические формулы. Локальное сглаживание данных. 83

Глава 1
ПРАВИЛА ПРИБЛИЖЕННЫХ ВЫЧИСЛЕНИЙ И ОЦЕНКА ПОГРЕШНОСТЕЙ ПРИ ВЫЧИСЛЕНИЯХ

§ 1. Приближенные числа, их абсолютные и относительные погрешности

Расчеты, как правило, производятся с приближенными значениями величин — *приближенными числами.* Уже исходные данные для расчета обычно даются с некоторыми погрешностями; в процессе расчета еще накапливаются погрешности от округления, от применения приближенных формул и т. п. Разумная оценка погрешности при вычислениях позволяет указать оптимальное количество знаков, которые следует сохранять при расчетах, а также в окончательном результате.

Погрешность приближенного числа *a,* т. е. разность $a - a_0$ между ним и точным значением a_0, обычно неизвестна.

Под *оценкой погрешности* приближенного числа *a* понимают установление неравенства вида

$$|a - a_0| \le \Delta_a \tag{1.1}$$

Число Δ_a называется *абсолютной погрешностью приближенного числа a* (иногда употребляют термин «предельная абсолютная погрешность»). Это число определяется неоднозначно: его можно увеличить. Обычно стараются указать, возможно, меньшее число Δ_a, удовлетворяющее неравенству (1.1).

Абсолютные погрешности записывают не более чем с двумя-тремя значащими цифрами (при подсчете числа значащих цифр не учитывают нулей, стоящих слева; например, в числе 0,010030 имеется 5 значащих цифр). В приближенном числе *a* не следует сохранять те разряды, которые подвергаются округлению в его абсолютной погрешности Δ_a.

ПРИМЕР 1.1. Длина и ширина комнаты, измеренные с точностью до 1 *см,* равны a = 5,43 *м* и b = 3,82 *м.* Оценить погрешность в определении площади комнаты S = *ab* = 20,7426 м^2.

Решение. По условию задачи Δ_a = 0,01 *м,* 0, Δ_b = 0,01 *м.*

Крайние возможные значения площади равны

$$(a + 0{,}01)\,(b + 0{,}01) = 20{,}8352\ м^2,$$
$$(a - 0{,}01)\,(b - 0{,}01) = 20{,}6502\ м^2;$$

сравнивая их с подсчитанным выше значением *S,* получаем оценку

$$|S - S_0| \le 0{,}0926,$$

что дает возможность указать абсолютную погрешность числа S в виде Δ_S = 0,0926 *м*^{2}.

Здесь разумно округлить значение Δ_S, например, так: Δ_S =0,093 *м*^{2} при Δ_S =0,10 *м*^{2} (абсолютные погрешности округляют в большую сторону). При этом приближенное значение площади можно записать в виде *S = 20,743 м*^{2}, или S = 20,74 *м*^{2}, или даже S = 20,7 *м*^{2}.

ПРИМЕР 1.2. В некоторую вычислительную машину мы можем ввести числа только с тремя значащими цифрами. С какой точностью мы можем ввести в нее числа π и 1/3?

Решение. Полагаем $\pi \approx 3{,}14 = a$ вместо π = 3,141592..., погрешность числа *a* можно оценить числом Δ_a =0,0016. Полагаем $1/3 \approx 0{,}333 = b$; погрешность числа *b* можно оценить числом Δ_b =0,00034 или Δ_b =0,0004.

Относительной погрешностью $\delta_S = \dfrac{0{,}0926}{20{,}7426} = \dfrac{926}{207426} = 0{,}0045 = 0{,}45\%.$ *приближенного числа a* называется отношение его абсолютной погрешности Δ_a к абсолютной величине числа *a,* т. е.

$$\frac{\Delta_a}{|a|} \tag{1.2}$$

$(a \ne 0)$. Относительная погрешность обычно выражается в процентах, и ее принято записывать не более чем с двумя-тремя значащими цифрами (знаками).

Иногда под относительной погрешностью понимают отношение, $\dfrac{\Delta_a}{|a_0|}$, где a_0 — точное (но неизвестное) значение числа; если относительная погрешность числа *a* не превышает 5%, то различие между

отношениями $\frac{\Delta_a}{|a|}$ и $\frac{\Delta_a}{|a_0|}$ сказывается только на втором знаке погрешности, что не существенно.

П р и м е р 1.3. Определить относительную погрешность числа S в примере 1.1.

Р е ш е н и е . S = 20,7426, Δ_S = 0,0926, поэтому

$$\delta_S = \frac{0,0926}{20,7426} = \frac{926}{207426} = 0,0045 = 0,45\%.$$

Во многих технических приложениях принято характеризовать точность приближенных чисел их относительной погрешностью.

Относительная погрешность приближенного числа связана с количеством его верных знаков. *Количество верных знаков* числа отсчитывается от первой значащей цифры числа до первой значащей цифры его абсолютной погрешности: например, число S = 20,7426 с абсолютной погрешностью Δ_S = 0,0926 имеет три верных знака (2, 0, 7); остальные знаки — сомнительные.

Ориентировочно можно считать, что наличие только одного верного знака соответствует относительной погрешности порядка 10%, двух верных знаков — погрешности порядка 1%, трех верных знаков — погрешности порядка 0,1% и т. д.

В математических таблицах все числа округлены до верных знаков, причем абсолютная погрешность не превосходит половины единицы последнего оставленного разряда. Например, если в таблице указано e=2,718, то абсолютная погрешность не превосходит $0,5\cdot10^{-3}$.

В окончательных результатах вычислений обычно оставляют, кроме верных, один сомнительный знак.

В промежуточных результатах вычислений обычно сохраняют два-три сомнительных знака, чтобы не накапливать лишних погрешностей от округлений.

П р и м е р 1.4. Округлить число S = 20,7426 в примере 1.1 до верных знаков.

Р е ш е н и е . Так как в числе S три верных знака, то естественно записать S=20,7.

Однако при этом к абсолютной погрешности Δ_S=0,0926 приходится добавить еще величину 0,0426, отброшенную при округлении. Новая абсолютная погрешность Δ_S = 0,136 заставляет считать сомнительным уже третий знак числа *S,* и, следовательно, число 5 приходится округлять до двух знаков:

S=21 (Δ_S= 0,44 < 0,5).

Этот пример показывает, что округление результатов расчета до верных знаков не всегда целесообразно.

Примечание. В этом примере, как это обычно принято, применено *правило дополнения при округлении:* если первая отбрасываемая цифра больше или равна 5, то последняя сохраняемая цифра увеличивается на 1.

ЗАДАЧИ

1. Округляя следующие числа до трех значащих цифр, определить абсолютную Δ и относительную δ погрешности полученных приближенных чисел.

а) 2,1514, б) 0,16152, в) 0,01204, г) 1,225, д) -0,0015281, е) -392,85, ж) 0,1545, з) 0,003922, и) 625,55, к) 94,525.

2. Определить абсолютную погрешность следующих приближенных чисел по их относительным погрешностям.

а) a=13267, δ=0,1%, б) a=2,32, δ = 0,7%, в) a = 35,72, δ = 1%, г) a =0,896, δ = 1 0 %, д) a= 232,44, δ =1%.

3. При измерении некоторых углов получили числа α_1 = 21°37'3", α_2 = 45°, α_3=1°10", α_4 = 75°20'44".

Определить относительные погрешности чисел $\alpha_1, \alpha_2, \alpha_3, \alpha_4$, полагая абсолютную погрешность измерения равной 1".

4. Определить количество верных знаков в числе *x,* если известна его абсолютная погрешность.

а) $x = 0,3941$, $\Delta_x = 0,25\cdot10^{-2}$ б) x = 0,1132, $\Delta_x = 0,1\cdot10^{-3}$, в) $\Delta_x = 38,2543$, $\Delta_x = 0,27\cdot10^{-2}$,

г) x = 293,481, Δ_x = 0,1, д) x = 2,325, $\Delta_x = 0,1\cdot10^{-1}$, е) x = 14,00231, $\Delta_x = 0,1\cdot10^{-3}$,

ж) x = 0,0842, $\Delta_x = 0,15\cdot10^{-2}$, з) x = 0,00381, $\Delta_x = 0,1\cdot10^{-4}$, и) $x = -32,285$, $\Delta_x = 0,2\cdot10^{-2}$,

к) $x = -0,2113$, $\Delta_x = 0,5\cdot10^{-2}$.

5.Определить количество верных знаков в числе *a,* если известна его относительная погрешность.

а) $a = 1,8921$, $\delta_a = 0,1\cdot10^{-2}$ б) $a = 0,2218$, $\delta_a = 0,2\cdot10^{-1}$, в) $a = 22,351$, $\delta_a = 0,15$,

г) $a = 0,02425$, $\delta_a = 0,5\cdot10^{-2}$, д) $a = 0,000135$, $\delta_a = 0,15$, е) $a = 9,3598$, $\delta_a = 0,1\%$,

ж) $a = 0,11452$, $\delta_a = 10\%$, з) $a = 48361$, $\delta_a = 1\%$, и) $a = 592,8$, $\delta_a = 2\%$, к) $a = 14,9360$, $\delta_a = 1\%$.

§ 2. Устойчивость. Корректность. Сходимость

1. Устойчивость. Рассмотрим погрешности исходных данных. Поскольку это так называемые неустранимые погрешности и вычислитель не может с ними бороться, то нужно хотя бы иметь представление об их влиянии на точность окончательных результатов. Конечно, мы вправе надеяться на то, что погрешность результатов имеет порядок погрешности исходных данных. Всегда ли это так? К сожалению, нет. Некоторые задачи весьма чувствительны к неточностям в исходных данных. Эта чувствительность характеризуется так называемой устойчивостью.

Пусть в результате решения задачи по исходному значению величины x находится значение искомой величины y. Если исходная величина имеет абсолютную погрешность Δx, то решение имеет погрешность Δy. Задача называется *устойчивой* но исходному параметру x, если решение y непрерывно от него зависит, т. е. малое приращение исходной величины Δx приводит к малому приращению искомой величины Δy. Другими словами, малые погрешности в исходной величине приводят к малым погрешностям в результате расчетов.

Отсутствие устойчивости означает, что даже незначительные погрешности в исходных данных приводят к большим погрешностям в решении или вовсе к неверному результату. О подобных неустойчивых задачах также говорят, что они *чувствительны* к погрешностям исходных данных.

Примером такой задачи является отыскание действительных корней многочленов вида

$$(x-a)^n = \varepsilon,\ \ 0 < \varepsilon << 1.$$

Изменение правой части па величину порядка ε приводит к погрешности корней порядка $\varepsilon^{1/n}$.

Интересной иллюстрацией неустойчивой задачи является так называемый *пример Уилкинсона.* Рассматривается многочлен

$$P(x) = (x-1)(x-2)...(x-20) = x^{20} - 210x^{19} + ...$$

Очевидно, что корнями этого многочлена являются $x_1 = 1, x_2 = 2, ..., x_{20} = 20$.

Предположим, что один из коэффициентов многочлена вычислен с некоторой малой погрешностью. Например, коэффициент -210 при x^{19} увеличим на 2^{-23} (около 10^{-7}). В результате вычислений даже с точностью до 11 значащих цифр получим существенно другие значения корней. Приведем для наглядности эти значения, округленные до трех знаков:

$x_1 = 1{,}00,$	$x_9 = 8{,}92,$
$x_2 = 2{,}00,$	$x_{10,11} = 10{,}1 \pm 0{,}644i,$
$x_3 = 3{,}00,$	$x_{12,13} = 11{,}8 \pm 1{,}65i,$
$x_4 = 4{,}00,$	$x_{14,15} = 14{,}0 \pm 2{,}52i,$
$x_5 = 5{,}00,$	$x_{16,17} = 16{,}7 \pm 2{,}81i,$
$x_6 = 6{,}00,$	$x_{18,19} = 19{,}5 \pm 1{,}94i,$
$x_7 = 7{,}00,$	$x_{20} = 20{,}8$
$x_8 = 7{,}01,$	

Таким образом, изменение коэффициента -210 при x^{i9} на -210 $+10^{-7}$ привело к тому, что половина корней стали комплексными. Причина такого явления — неустойчивость самой задачи; вычисления выполнялись очень точно (11 разрядов), а погрешности округлений не могли привести к таким последствиям.

2. Корректность. Задача называется *поставленной корректно,* если для любых значений исходных данных из некоторого класса ее решение существует, единственно и устойчиво по исходным данным.

Рассмотренные выше примеры неустойчивых задач являются некорректно поставленными. Применять для решения таких задач численные методы, как правило, нецелесообразно, поскольку возникающие в расчетах погрешности округлений будут сильно возрастать в ходе вычислений, что приведет к значительному искажению результатов.

Вместе с тем отметим, что в настоящее время развиты методы решения некоторых некорректных задач. Это в основном так называемые *методы регуляризации.* Они основываются на замене исходной задачи корректно поставленной задачей. Последняя содержит некоторый параметр, при стремлении которого к нулю решение этой задачи переходит в решение исходной задачи.

3. Неустойчивость методов. Иногда при решении корректно поставленной задачи может оказаться неустойчивым метод ее решения. В частности, по этой причине при вычислении синуса большого аргумента был получен результат, не имеющий смысла.

Рассмотрим еще один пример неустойчивого алгоритма. Достроим численный метод вычисления интеграла

$$I_n = \int_0^1 x^n e^{x-1} dx\ ,\ \text{n} = 1, 2, \ldots$$

Интегрируя по частям, находим

$$I_1 = \int_0^1 xe^{x-1}dx = xe^{x-1}\Big|_0^1 - \int_0^1 e^{x-1}dx = \frac{1}{e}$$

$$I_2 = \int_0^1 x^2e^{x-1}dx = x^2e^{x-1}\Big|_0^1 - 2\int_0^1 xe^{x-1}dx = 1-2I_{n-1}$$

……

$$I_2 = \int_0^1 x^ne^{x-1}dx = x^ne^{x-1}\Big|_0^1 - n\int_0^1 x^{n-1}e^{x-1}dx = 1-nI_1 \quad .$$

Пользуясь полученным рекуррентным соотношением, вычисляем

I_1 = 367879, I_6 = 0,127120,
I_2 = 0,263242, I_7 = 0,110160,
Iз = 0,207274, I_8 = 0,118720,
I_4 = 0,170904, I_9 = -0,0684800.
I_5 = 0,145480,

Значение интеграла I_9 не может быть отрицательным, поскольку подынтегральная функция на всем отрезке интегрирования [0, 1] неотрицательна. Исследуем источник погрешности. Видим, что округление в I_1 дает погрешность, равную примерно лишь $4.4*10^{-7}$. Однако на каждом этапе эта погрешность умножается на число, модуль которого больше единицы (-2, -3, ..., -9), что в итоге дает 9. Это и приводит к результату, не имеющему смысла. Здесь снова причиной накопления погрешностей является алгоритм решения задачи, который оказался неустойчивым.

Численный алгоритм (метод) называется *корректным* в случае существования и единственности численного решения при любых значениях исходных данных, а также в случае устойчивости этого решения относительно погрешностей исходных данных.

4. Понятие сходимости. При анализе точности вычислительного процесса одним из важнейших, критериев является *сходимость* численного метода. Она означает близость получаемого численного решения задачи к истинному решению. Строгие определения разных оценок близости могут быть даны лишь с привлечением аппарата функционального анализа. Здесь мы ограничимся некоторыми понятиями сходимости, необходимыми для понимания последующего материала.

Рассмотрим понятие *сходимости итерационного процесса.* Этот процесс состоит в том, что для решения некоторой задачи и нахождения искомого значения определяемого параметра (например, корня нелинейного уравнения) строится метод последовательных приближений. В результате многократного повторения этого процесса (или *итераций)* получаем последовательность значений x_1, x_2, ..., x_n. Говорят, что эта последовательность *сходится* к точному решению $x = a$, если при неограниченном возрастании числа итераций предел этой последовательности существует и равен a: $\lim\limits_{n\to\infty} x_n = a$. В этом случае имеем сходящийся численный метод.

Другой подход к понятию сходимости используется в методах дискретизации. Эти методы заключаются в замене задачи с непрерывными параметрами на задачу, в которой значения функций вычисляются в фиксированных точках. Это относится, в частности, к численному интегрированию, решению дифференциальных уравнений и т. п. Здесь под *сходимостью метода* понимается стремление значений решения дискретной модели задачи к соответствующим значениям решения исходной задачи при стремлении к нулю параметра дискретизации (например, шага интегрирования).

При рассмотрении сходимости важными понятиями являются ее вид, порядок и другие характеристики. С общей точки зрения эти понятия рассматривать нецелесообразно; к ним будем обращаться при изучении численных методов.

Таким образом, для получения решения задачи с необходимой точностью ее постановка должна быть корректной, а используемый численный метод должен обладать устойчивостью и сходимостью.

§ 3. Сложение и вычитание приближенных чисел

1. Абсолютная погрешность алгебраической суммы нескольких приближенных чисел равна сумме абсолютных погрешностей слагаемых: если

$$S=a_1+a_2+\ldots+a_n, \text{ то}$$

$$\Delta_S = \Delta_{a_1} + \Delta_{a_2} + \ldots + \Delta_{a_n}. \qquad (1.3)$$

При большом количестве слагаемых оценка абсолютной погрешности суммы по формуле (1.3) оказывается сильно завышенной, так как обычно происходит частичная компенсация погрешностей разных знаков.

Если среди слагаемых имеется одно число, абсолютная погрешность которого значительно превосходит абсолютные погрешности остальных слагаемых, то абсолютная погрешность суммы считается равной этой наибольшей погрешности. При этом в сумме целесообразно сохранять столько десятичных знаков, сколько их в слагаемом с наибольшей абсолютной погрешностью.
Покажем на примере, как производится сложение приближенных чисел и оценка погрешности.

ПРИМЕР 1.5. Найти сумму приближенных чисел 0,348; 0,1834; 345,4; 235,2; 11,75; 9,27; 0,0849; 0,0214; 0,000354, считая в них все знаки верными, т. е. считая, что абсолютная погрешность каждого слагаемого не превосходит половины единицы младшего оставленного разряда.
Решение. Наибольшую абсолютную погрешность $\Delta = 0{,}05$ имеют два числа: 345,4 и 235,2. Поэтому можно считать, что абсолютная погрешность суммы составляет $2\Delta = 0{,}10$. Так как количество слагаемых невелико, то в расчетах сохраняем только один запасной знак, т. е. округляем слагаемые до 0,01:

345,4
235,2
11,75
9,27
0,35
0,18
0,08
0,02
0,00
S = 602,25.

В окончательном результате запасной знак отбрасываем:

S = 602,2

При этом к указанной выше абсолютной погрешности 0,10 добавляем погрешность округления 0,05, что дает

$$\Delta_S = 0{,}15 \text{ или } \Delta_S = 0{,}2.$$

Заметим, что в этом примере полный учет всех погрешностей слагаемых по формуле (1.3) только усложнил бы расчет, не внося существенных уточнений в результат.
2. Относительная погрешность δ_S суммы нескольких чисел одного и того же знака заключена между наименьшей и наибольшей из относительных погрешностей слагаемых:

$$\min \delta_{a_k} \le \delta_S \le \max \delta_{a_k} \ (a_k > 0,\ k = 1, 2, \ldots, n) \qquad (1.4)$$

ПРИМЕР 1.6. Оценить относительную погрешность суммы чисел в примере 1.5 и сравнить ее с относительными погрешностями слагаемых.
Решение. Относительная погрешность суммы S равна

$$\delta_S = \frac{0{,}10}{602{,}25} = 0{,}017\%$$

Относительные погрешности слагаемых составляют соответственно

$$\frac{0{,}5}{348} = 0{,}15\%;\quad \frac{0{,}5}{1834} = 0{,}027\%;\quad \frac{0{,}5}{3454} = 0{,}015\%$$

$$\frac{0{,}5}{2352} = 0{,}022\%;\quad \frac{0{,}5}{1175} = 0{,}043\%;\quad \frac{0{,}5}{927} = 0{,}054\%;$$

$$\frac{0{,}5}{849} = 0{,}059\%;\quad \frac{0{,}5}{214} = 0{,}24\%;\quad \frac{0{,}5}{354} = 0{,}15\%$$

ЗАДАЧИ

1. Найти суммы приближенных чисел и указать их погрешность
а) 0,145 + 321 +78,2 (все знаки верные),
б) 0,301 + 193,1 + 11,58 (все знаки верные),
в) 398,5 - 72,28 + 0,34567 (все знаки верные),
г) $x_1 + x_2 - x_3$, где $x_1 = 197{,}6$, $\Delta_{x_1} = 0{,}2$, $x_2 = 23{,}44$, $\Delta_{x_2} = 0{,}22$, $x_3 = 201{,}55$, $\Delta_{x_3} = 0{,}17$.

§ 4. Умножение и деление приближенных чисел

При умножении и делении приближенных чисел их относительные погрешности складываются (а не абсолютные). Относительная погрешность выражения

$$r = \frac{a_1 a_2 \ldots a_m}{b_1 b_2 \ldots b_n} \qquad (1.5)$$

Оценивается величиной

$$\delta_r = \delta_{a_1} + \delta_{a_2} + \ldots + \delta_{a_m} + \delta_{b_1} + \delta_{b_2} + \ldots + \delta_{b_n} \qquad (1.6)$$

При большом числе $m+n$ выгоднее пользоваться *статистической оценкой,* учитывающей частичную компенсацию погрешностей разных знаков: если все числа a_i, b_j ($i = 1, 2, \ldots, m;\ j=1,2, \ldots, n$) имеют примерно одинаковую относительную погрешность δ, то относительная погрешность выражения (1.5) принимается равной

$$\delta_r = \sqrt{3(n+m)}\,\delta \qquad (n+m>10) \qquad (1.7)$$

Если у одного из чисел !!!a_h bj *!!!* относительная погрешность значительно превышает относительные погрешности остальных чисел, то относительная погрешность выражения (1.3) считается равной этой наибольшей погрешности. При этом в результате целесообразно сохранять столько знаков (значащих цифр), сколько их в числе с наибольшей относительной погрешностью.

Абсолютная погрешность выражения (1.5) вычисляется по его относительной погрешности:

$$\Delta_r = |r|\,\delta_r .$$

Покажем на примере, как производится умножение и деление приближенных чисел и *оценка* погрешности результата.

П р и м е р 1.7. Вычислить выражение

$$r = \frac{3,2 \cdot 356,7 \cdot 0,04811}{7,1948 \cdot 34,56},$$

считая, что все числа даны с верными знаками, т. е. что их абсолютные погрешности не превосходят половины единицы младшего оставленного разряда.

Р е ш е н и е. Наибольшую относительную погрешность имеет число а = 3,2, которое содержит всего два верных знака (против четырех-пяти верных знаков в остальных числах):

$$\delta_a = \frac{0,5}{32} = 1,6\% .$$

Поэтому можно считать, что относительная погрешность результата составляет δ_a= 1,6%, т. е. что результат содержит не более двух верных знаков. Так как количество'данных чисел невелико, то в расчетах сохраняем один запасной знак, округляя все числа до трех знаков:

$$r = \frac{3,2 \cdot 357 \cdot 0,0481}{7,19 \cdot 34,6} = 0,221$$

Абсолютную погрешность результата вычисляем по его относительной погрешности и найденному численному значению:

$$\Delta_r = r\delta_r = 0,221 \cdot 0,016 = 0,0036 .$$

Округляя результат до верных знаков, отбрасываем запасной знак и получаем

$$r = 0,22$$

с абсолютной погрешностью $\Delta_r < 0,005$.

З А Д А Ч И

1. Найти произведения приближенных чисел и указать их погрешности (считая в исходных данных все знаки верными).

а) 3,49*8,6, б) 25,1*1,743, в) 0,02*16,5, г) 0,253*654*83,6, д) 1,78*9,1*1,183, е) 482,56*7256*0,0052.

2. Найти частное приближенных чисел.

а) 5,684/5,032, б) 0,144/1,2, в) 216/4, г) 726,676/829, д) 754,9367/36,5, е) 7,3/4491.

Стороны прямоугольника равны (4,02 ± 0,01) *м,* (4,96 ± 0,01) *м.* Вычислить площадь прямоугольника.

Катеты, прямоугольного треугольника равны (12,10 + 0,01) *см,* (25,21 + 0,01) *см.* Вычислить тангенс угла, противолежащего первому катету.

При измерении радиуса *R* круга с точностью до 0,5 *см* получилось число 12 *см.* Найти абсолютную и относительную погрешности при вычислении площади круга.

Каждое ребро куба, измеренное с точностью до 0,02 *см,* оказалось равным 8 *см.* Найти абсолютную и относительную погрешности при вычислении объема куба.

Высота *h* и радиус основания *R* цилиндра измерены с точностью до 0,5%. Какова предельная относительная погрешность при вычислении объема цилиндра?

§ 5. Погрешности вычисления значений функции

1. Функции одной переменной. Абсолютная погрешность дифференцируемой функции $y = f(x)$, вызываемая достаточно малой погрешностью аргумента A_x, оценивается величиной

$$\Delta_y = \left| f'(x) \right| \Delta_x \tag{1.8}$$

Если значения функции *f(x)* положительны, то для относительной погрешности имеет место оценка

$$\delta_y = \frac{\left| f'(x) \right|}{f(x)} \Delta_x = \left| \left[\ln f(x) \right]' \right| \Delta_x \tag{1.9}$$

В частности, для основных элементарных функций получаем следующие правила.

а) Степенная функция $y = x^a$. Абсолютная погрешность степенной функции равна

$$\Delta_y = a x^{a-1} \Delta_x \tag{1.10}$$

Относительная погрешность степенной функции равна

$$\delta_y = \left| a \right| \delta_x \tag{1.11}$$

Например, относительная погрешность квадрата x^2 вдвое больше относительной погрешности основания x, относительная погрешность квадратного корня $\sqrt{x}$ вдвое меньше относительной погрешности подкоренного числа *x,* относительная погрешность обратной величины 1/x равна относительной погрешности самого числа x.

б) Показательная функция $y = a^x$ $(a > 0)$. Абсолютная погрешность показательной функции равна

$$\Delta_y = a^x \ln a \cdot \Delta_x \tag{1.12}$$

Относительная погрешность показательной функции равна

$$\delta_y = \Delta_x \ln a \tag{1.13}$$

Заметим, что здесь относительная погрешность функции пропорциональна абсолютной погрешности аргумента. Для функции $y = e^x$ отсюда получаем

$$\delta_y = \Delta_x \tag{1.14}$$

в) Логарифмическая функция $y = \ln x$. Абсолютная погрешность натурального логарифма числа равна относительной погрешности самого числа:

$$\Delta_y = \frac{1}{x} \Delta_x = \delta_x \tag{1.15}$$

Для десятичного логарифма $y = \lg x$ имеем

$$\Delta_y = 0,4343 \delta_x \tag{1.16}$$

откуда следует, что при расчетах с числами, имеющими *m* верных знаков, надо пользоваться *(m*+1)-значными таблицами логарифмов.

г) Тригонометрические функции. Абсолютные погрешности синуса и косинуса не превосходят абсолютных погрешностей аргумента:

$$\Delta_{\sin x} = \left| \cos x \right| \Delta_x \le \Delta_x \qquad \Delta_{\cos x} = \left| \sin x \right| \Delta_x \le \Delta_x \tag{1.17}$$

Абсолютная погрешность тангенса и котангенса всегда больше абсолютной погрешности аргумента:

$$\Delta_{tgx} = (1 + tg^2 x) \Delta_x \ge \Delta_x, \quad \Delta_{ctgx} = (1 + ctg^2 x) \Delta_x \ge \Delta_x \tag{1.18}$$

П р и м е р 1.8. Диаметр круга, измеренный с точностью до 1 *мм,* оказался равным *d* = 0,842 *м.* Вычислить площадь круга.

Р е ш е н и е . Площадь круга $S = \pi d^2 / 4$. Так как число π мы можем взять для расчета с любой точностью, то погрешность вычисления площади определяется погрешностью вычисления. Относительная погрешность d^2 равна

$$\delta_{d^2} = 2\delta_d = 2 \cdot \frac{1}{842} = 0,24\%$$

Чтобы при округлении числа π не увеличить относительную погрешность

$$\delta_S = \delta\left(\frac{\pi}{4} \right) + 2\delta_d ,$$

надо взять число π по крайней мере с четырьмя верными **знаками,** еще лучше с пятью. Тогда получим

$$S = \frac{3,1416}{4} \cdot 0,842^2 \text{ м}^2 = 0,7854\text{-}0,7090 \text{ м}^2 = 0,5568 \text{ м}^2.$$

Абсолютная погрешность результата составляет

$$\Delta_S = S\delta_S = 0,557 \cdot 0,0024 = 0,0014 .$$

Округляем результат до трех знаков (отбрасывая запасной знак и пользуясь правилом дополнения):

$$S = 0{,}557\ м^2, \qquad \Delta_S = 0{,}002.$$

ПРИМЕР 1.9. Угол x = 25°20' измерен с точностью до 1'. Определить $\sin x$ и его абсолютную погрешность.

Решение. Вычислим сначала абсолютную погрешность $\sin x$ по формуле (1.17) для этого надо еще перевести 1' в радианы: 1'=0,000291 — и подсчитать

$$\Delta_{\sin x} = \cos x \cdot \Delta_x = \cos 25^\circ 20' = 0{,}4279$$

Поэтому для вычисления *sin(x)* надо взять четырехзначные таблицы тригонометрии ческих функций, что дает $\sin x = \sin 25°20' = 0{,}4279$.

2. **Функции нескольких переменных.** Абсолютная погрешность дифференцируемой функции $y = f(x_1, x_2, \dots x_n)$, вызываемая достаточно малыми погрешностями Δ_{x_1}, Δ_{x_2}, ..., Δ_{x_n} аргументов x_1, x_2, ..., x_n, оценивается величиной

$$\sqrt{y} = \sqrt{4{,}4} = 2{,}10 \tag{1.19}$$

Если значения функции положительны, то для относительной погрешности имеет место оценка

$$\delta_y = \sum_{i=1}^{n} \frac{1}{f}\left|\frac{\partial f}{\partial x_i}\right| \Delta_{x_i} = \sum_{i=1}^{n} \left|\frac{\partial \ln f}{\partial x_1}\right| \tag{1.20}$$

ПРИМЕР 1.10. Вычислить значение функции $u = xy^2z^3$, если $x = 37{,}1$, $y = 9{,}87$, $z = 6{,}052$, причем $\Delta_x = 0{,}3$, $\Delta_y = 0{,}11$, $\Delta_z = 0{,}016$.

Решение. Здесь относительные погрешности аргументов равны

$$\delta_x = \frac{3}{371} = 0{,}81\%, \qquad \delta_y = \frac{11}{987} = 1{,}12\%, \qquad \delta_z = \frac{16}{6052} = 0{,}26\%.$$

Относительная погреишость функции равна

$$\delta_u = \delta_x + 2\delta_y + 3\delta_z = 3{,}8\%;$$

поэтому значение функции следует вычислять не более чем с двумя-тремя знаками:

$$u = 801 \cdot 10^3$$

(нельзя писать 801 000, это имело бы другой смысл). Абсолютная погрешность при этом составляет

$$\Delta_u = u\delta_u = 801 \cdot 10^3 \cdot 0{,}038 = 30 \cdot 10^3.$$

Здесь целесообразно округлить результат до двух знаков:

$$u = 8{,}0 \cdot 10^5, \qquad \Delta_u = 0{,}3 \cdot 10^5.$$

ПРИМЕР 1.11. Вычислить значение $z = \ln(10{,}3 + \sqrt{4{,}4})$, считая верными все знаки приближенных чисел x=10,3 и y==4,4.

Решение. Число y имеет относительную погрешность $\delta_y = \frac{0{,}5}{44} = 1{,}2\%$, поэтому $\sqrt{y}$ имеет относительную погрешность 0,6% и его следует записать с тремя знаками:

$$\sqrt{y} = \sqrt{4{,}4} = 2{,}10;$$

при этом абсолютная погрешность этого корня равна $\Delta_{\sqrt{y}} = 2{,}10 \cdot 0{,}006 = 0{,}013$.

Абсолютная погрешность суммы $x + \sqrt{y} = 10{,}3 + 2{,}10 = 12{,}4$ оценивается величиной 0,05 + 0,013 = 0,063, ее относительная погрешность равна $\frac{0{,}63}{124} = 0{,}5\%$.

По формуле (1.15) такова же будет абсолютная погрешность натурального логарифма, т. е. $\Delta_z = 0{,}005$. Поэтому $z = \ln(10{,}3 + 2{,}10) = \ln 12{,}40 = 2{,}517$.

Здесь результат имеет три верных знака; округление до верных знаков нецелесообразно, так как при этом надо писать значение Δ_z с учетом погрешности округления:

$$z = 2{,}52, \qquad \Delta_z = 0{,}008.$$

ЗАДАЧИ

1. Углы x измерены с предельной абсолютной погрешностью Δ_x.Определить абсолютную и относительную

погрешности функций $y = \sin x$, $y = \cos x$ и $y = tgx$. Найти по таблицам значения функций, сохранив в результате лишь верные цифры.

а) $x = 11^\circ 20'$, $\Delta_x = 1'$, б) x = 48°42'31", $\Delta x = 5"$, в) x = 45°, $\Delta_x = 1'$, г) x = 50°10', $\Delta x = 0{,}05°$,

д) x = 0,45, $\Delta x = 0{,}5 \cdot 10^{-2}$, е) x = 1,115, $\Delta x = 0{,}1 \cdot 10^{-3}$.

2. Для следующих функций вычислить значения при указанных значениях x и указать абсолютную и относительную погрешности результатов.

а) $y = x^3 \sin x$ при $x = \sqrt{2}$, полагая $\sqrt{2} \approx 1{,}414$,

б) $y = x \ln x$ при $x = \pi$, полагая $\pi \approx 3{,}142$,

в) $y = e^x \cos x$ при $x = \sqrt{3}$, полагая $\sqrt{3} \approx 1{,}732$.

3. Для следующих функций вычислить значения при указанных значениях переменных. Указать абсолютную и относительную по грешности результатов, считая все знаки исходных данных верными.

а) $u = \ln(x_1 + x_2^2)$, $x_1 = 0{,}97$, $x_2 = 1{,}132$,

б) $u = \dfrac{x_1 + x_2^2}{x_3}$, $x_1 = 3{,}28$, $x_2 = 0{,}932$, $x_3 = 1{,}132$,

в) $u = x_1x_2 + x_1x_3 + x_2x_3$, $x_1 = 2{,}104$, $x_2 = 1{,}935$, $x_3 = 0{,}845$.

4. Определить относительную погрешность при вычислении полной поверхности усеченного конуса, если радиусы его оснований R и r и образующая l, измеренные с точностью до 0,01 *см,* равны

R =23,64 *см,* r =17,31 *см,* l =10,21 *см.*

§ 6. Определение допустимой погрешности аргументов по допустимой погрешности функции

Эта задача имеет однозначное решение только для функции одной переменной $y = f(x)$: если эта функция дифференцируема и $f'(x) \neq 0$, то

$$\Delta_x = \frac{1}{|f'(x)|}\Delta_y \tag{1.21}$$

Для функции нескольких переменных $y=f(x_1, x_2, ..., x_n)$ задача решается только при введении каких-либо дополнительных ограничений. Например, если значение одного из аргументов значительно труднее измерить или вычислить с большой точностью, чем значения остальных аргументов, то погрешность именно этого аргумента надо согласовать с требуемой погрешностью функции.

Если значения всех аргументов можно одинаково легко определить с любой точностью, то обычно применяют *принцип равных влияний,* считая, что в формуле (4.12)!!!(в источнике ссылается на формулу, которая 1.19)!!!! все слагаемые $\left|\frac{\partial f}{dx_i}\right|\Delta_{x_i}$ равны между собой; это дает формулу

$$\Delta_{x_i} = \frac{\Delta_y}{n\left|\frac{\partial f}{dx_i}\right|} \qquad (i = 1, 2, \ ..., n) \tag{1.22}$$

На практике часто встречаются задачи промежуточного типа между указанными крайними случаями. Мы рассмотрим соответствующие примеры.

ПРИМЕР 1.12. С какой точностью следует измерить угол x в первой четверти, чтобы получить значение sinx с пятью верными знаками?

Решение. Если известно, что угол $x > 6°$, так что $\sin x > 0{,}1$, то надо определить Δ_x так, чтобы выполнялось неравенство $\Delta_{\sin x} < 0{,}5 \cdot 10^{-5}$. Для этого в соответствии с формулой (1.17) достаточно взять $\Delta_x < 0{,}5 \cdot 10^{-5}$, т. е. измерить угол x с точностью до 1". Если, сверх того, известно, что угол $x > 60°$ и, значит, $\cos x < 0{,}5$, то стоит воспользоваться формулой (1.21), откуда $\Delta_x = \dfrac{1}{\cos x}\Delta_{\sin x} > 2 \cdot 0{,}5 \cdot 10^{-5} = 10^{-5}$,

т. е. достаточно измерить угол x с точностью всего до 2".

Но если угол $x < 6°$, например, $1° < x < 6°$, то $0{,}01 < \sin x < 0{,}1$ и для обеспечения пяти верных знаков в значении sin л: придется обеспечить неравенство $\Delta_{\sin x} < 0{,}5 \cdot 10^{-6}$, для чего придется измерять угол x с точностью до 0,1".

ПРИМЕР 1.13. С какой точностью следует определить радиус основания R и высоту H цилиндрической банки, чтобы ее вместимость можно было определить с точностью до 1%?

Р е ш е н и е . В формуле $V = \pi R^2 H$ число я можно взять с любым числом верных знаков, так что его погрешность не скажется на результате, и поэтому можно считать $\delta_V = 2\delta_R + \delta_H$. Если можно обеспечить любую точность определения *R* и *H,* то можно воспользоваться принципом равных влияний, откуда па долю $2\delta_R$ и δ_H приходится по 0,5%. Таким образом, но принципу равных влияний надо определить радиус с относительной погрешностью 0,25%, а высоту — с относительной погрешностью 0,5%. На практике чаще встречаются такие случаи, когда, наоборот, радиус банки определяется с меньшей точностью, чем высота. Например, если радиус определяется с точностью вдвое меньшей, чем высота, то полагаем $\delta_R = 2\delta_H$ и из условия

$$2\delta_R + \delta_H = 5\delta_H = 1\%$$

находим

$$\delta_H = 0,2\%, \quad \delta_R = 0,4\%.$$

Что касается числа π , то во всех указанных случаях надо брать его с относительной погрешностью порядка 0,01%, чтобы эту погрешность можно было не учитывать в окончательном результате. Это означает, что можно положить π = 3,142 с относительной погрешностью $\frac{0,4}{3142} = 0,013\%$, но не следует полагать π =3,14 с относительной погрешностью $\frac{0,16}{314} = 0,051\%$.

П Р И М Е Р 1.14. Найти допустимую абсолютную погрешность приближенных величин *x* = 15,2, *y* = 57°, для которых возможно найти значение функции

$$u = 6x^2(\lg x - \sin 2y)$$

с точностью до двух десятичных знаков (после запятой).

Р е ш е н и е . Находим u= $6x^2$ (lg x-sin *2y)* = 6 $(15,2)^2$ (lg 15,2 —sin 114°) = **371,9,**

$$\frac{du}{dx} = 12x(\lg x - \sin 2y) + 6x \cdot \lg e = 88,54$$

$$\frac{du}{dy} = -12x^2 \cos 2y = 1127,7$$

.

По условию $\Delta_u = 0,005$. Тогда согласно принципу **равных влияний** по формуле (1.22) находим

$$\Delta_x = \frac{\Delta_u}{2\left|\frac{du}{dx}\right|} = \frac{0,005}{2 \cdot 88,54} = 0,28 \cdot 10^{-4},$$

$$\Delta_y = \frac{\Delta_u}{2\left|\frac{du}{dy}\right|} = \frac{0,005}{2 \cdot 1127,7} = 0,22 \cdot 10^{-5} = 0'',45.$$

ЗАДАЧИ

1. С какой точностью следует взять приближенные числа x, чтобы значения sinx могли быть найдены с указанным числом *m* верных знаков?

а) x = 1°, *m* = 3, б) x = 25°, *m* = 4, в) x = 30,75°, *m=3,* г) x=1,05, *m* = 2, д) *x* = 0,075, *m* =2.

2. С какой точностью определены углы x по значениям sin *x,* взятым из пятизначной таблицы"функций?

а) *x* = 2°1', б) x = 15°30', в) x = 44°, г) x = 50°18', д) x = 65°23',

е) x = 87^0.

3. С какой точностью может быть определено число *x* по логарифму с помощью пятизначной таблицы логарифмов, если число находится в указанных пределах?

а) 300<x<400, б) 35<x<40, в) 1 , 5 < x < 1 , 7 , г) 3,25 < *x* < 3,29, д) 5000 < *x* < 6000.

4. С каким числом верных знаков следует взять значение аргумента *x,* чтобы получить значения указанных функций с точностью до 0,1-10^{-5}?

а) $y = x^3 \sin x$, $x = \sqrt{2}$, б) $y = x \ln x$, $x = \pi$, в) $y = e^x \cos x$, $x = \sqrt{3}$.

5. С каким числом верных знаков должен быть известен свободный член уравнения x^2 — 2x + lg2 = 0, чтобы получить корни с четырьмя верными знаками?

6. Найти допустимые абсолютные погрешности аргументов, которые позволяют вычислять значения данных

функций с четырьмя верными знаками.

а) $u = \ln(x_1 + x_2^2)$, $x_1 = 0,9731$, $x_2 = 1,13214$,

б) $u = \dfrac{x_1 + x_2^2}{x_3}$, $x_1 = 3,2835$, $x_2 = 0,93221$, $x_3 = 1,13214$,

в) $u = x_1x_2 + x_1x_3 + x_2x_3$, $x_1 = 2,10415$, $x_2 = 1,93521$, $x_3 = 0,84542$.

7. Для определения модуля Юнга по прогибу стержня прямоугольного сечения применяется формула

$$E = \frac{1}{4}\frac{l^3P}{d^3bs}$$

где l — длина стержня, *b* и *d* — основание и высота поперечного сечения, s— стрела прогиба, *P*—нагрузка. С какой точностью следует измерить длину l и стрелу s, чтобы погрешность E не превышала 5,5% при условии, что P известна с точностью до 0,1%, величины *b* и *d* известны с точностью до 1%, $l \approx 50$ *см,* $s \approx 2,5$ *см?*

Глава 2
ВЫЧИСЛЕНИЕ ЗНАЧЕНИЙ ФУНКЦИИ

При вычислении с помощью счётных машин значений функций, заданных формулами, далеко не безразлично, в каком виде записана соответствующая формула. Математически эквивалентные выражения часто оказываются неравноценными с точки зрения практики вычислений. Дело в том, что основными операциями большинства вычислительных машин являются сложение, вычитание, умножение и деление. Поэтому возникает необходимость представить рассматриваемую математическую задачу в виде последовательности этих элементарных операций. Учитывая ограниченность объёма памяти машины и необходимость экономии машинного времени, желательно эти операции разбить на повторяющиеся циклы и выбрать соответствующий алгоритм. Ниже мы рассмотрим приёмы, сводящие вычисление некоторых функций к таким циклам из элементарных операций.

§ 1. Вычисление значений многочлена. Схема Горнера

Пусть дан многочлен n-й степени

$$P(x) = a_0 x^n + a_1 x^{n-1} + \ldots + a_n$$

с действительными коэффициентами $a_k \ (k = 0,1,\ldots,n)$, и пусть требуется найти значение этого многочлена при $x = \xi$

$$P(\xi) = a_0 \xi^n + a_1 \xi^{n-1} + \ldots + a_n \tag{2.1}$$

Вычисление значения $P(\xi)$ удобнее всего производить следующим образом. Представим выражение $P(\xi) = a_0 \xi^n + a_1 \xi^{n-1} + \ldots + a_n$ (2.1) в виде

$$P(\xi) = \Big(\Big(\Big(\big(a_0 \xi + a_1\big)\xi + a_2\Big)\xi + a_3\Big)\xi + \ldots + a_n\Big).$$

Если ввести числа

$$\left.\begin{array}{ll} & b_0 = a_0, \\ c_1 = b_0\xi, & b_1 = a_1 + c_1, \\ c2 = b_1\xi, & b_2 = a_2 + c_2, \\ \ldots\ldots\ldots & \ldots\ldots\ldots \\ c_n = b_{n-1}\xi, & b_n = a_n + c_n, \end{array}\right\} \tag{2.2}$$

то $b_n = P(\xi).$

Таким образом, вычисление значения многочлена $P(x)$ при $x = \xi$ сводится к повторению следующей совокупности элементарных операций:

$$c_k = b_{k-1}\xi \quad b_k = a_k + c_k \quad (k = 1,2,\ldots,n).$$

Нетрудно показать, что числа $b_0 = a_0, b_1, \ldots, b_{n-1}$ являются коэффициентами многочлена $Q(x)$, полученного в качестве частного при делении данного многочлена $P(x)$ на двучлен $x - \xi$, а $b_n = P(\xi)$ – остаток от деления.

Таким образом, можно, не производя деления, определять коэффициенты частного $Q(x)$, а также остаток $P(\xi)$. Числа $b_0, b_1, \ldots, b_n$ обычно находят, пользуясь известной схемой *Горнера*

$$\begin{array}{ccccccc} a_0 & a_1 & a_2 & \ldots & a_n & \Big| & \xi \\ + & b_0\xi & b_1\xi & \ldots & b_{n-1}\xi & & \\ \hline b_0 = a_0 & b_1 & b_2 & \ldots & \underline{b_n = P(\xi)} & & \end{array}$$

Вычисление значений многочлена $P_n(x)$ по схеме Горнера требует выполнения n умножений и $n-k$ сложений, где k – число коэффициентов a_i, равных нулю. Если $a_0 = 1$, то требуется выполнить $n-1$ умножений. Показано, что для многочленов общего вида нельзя построить схему более экономную в смысле числа операций, чем схема Горнера.

П р и м ер 1.1. Вычислить при $x=-1,5$ значение многочлена

$$P(x)=x^7-2x^6+x^5-3x^4+4x^3-x^2+6x-1.$$

Р е ш е н и е. Пользуясь схемой Горнера, получим

	1	–2	1	–3	4	–1	6	–1	$\lfloor\xi$
+	–1,5	–1,5	5,25	–9,375	18,5625	–33,8438	52,2657	–87,3985	
	1	–3,5	6,25	–12,375	22,5625	–34,8438	58,2657	–88,3985	$=P(-1,5).$

Таким образом,

$$P(-1,5)=-88,3985.$$

З А Д А Ч И

1. Дан многочлен

$$P(x)=a_0x^4+a_1x^3+a_2x^2+a_3x+a_4.$$

Найти значение $P(3,25)$ для коэффициентов a_0, a_1, a_2, a_3, a_4, приведённых в Таблица 2.1

Таблица 2.1

	a_0	a_1	a_2	a_3	a_4		a_0	a_1	a_2	a_3	a_4
а)	7,54	11,08	3,82	0,44	-0,48	д)	2,79	9,85	14,15	5,38	7,24
б)	9,36	12,69	14,39	0,79	-094	е)	3,45	-2,91	3,79	-6,75	-2,38
в)	12,78	14,35	17,19	1,34	-1,72	ж)	4,79	5,38	-2,86	7,31	4,55
г)	15,65	17,58	21,7	2,78	1,34	з)	8,34	-7,75	4,53	-9,29	5,79

2. Дан многочлен

$$P(x)=0,22x^5-3,27x^4-2,74x^3-2,81x^2-3,36x+2.$$

Найти значения $P(\xi)$, где $\xi=0,80+0,05k;\ \ k=0,1,2,\ldots,20.$

§ 2. Вычисление значений некоторых трансцендентных функций с помощью степенных рядов

Здесь рассматриваются только такие трансцендентные функции, которые являются суммами своих рядов Маклорена

$$f(x)=\sum_{k=0}^{\infty}\frac{f^{(k)}(0)}{k!}x^k. \qquad (2.3)$$

Беря сумму нескольких первых членов ряда Маклорена, получаем приближённую формулу

$$f(x)\approx P_n(x),$$

$$P_n(x)=\sum_{k=0}^{\infty}\frac{f^{(k)}(0)}{k!}x^k.$$

При этом остаток ряда

$$R_n(x)=f(x)-P_n(x)$$

Представляет ошибку при замене $f(x)$ многочленом $P_n(x)$. Оценка остатка позволяет определить требуемое число слагаемых, т.е. степень n многочлена $P_n(x)$.

Заметим, что так как расчёт суммарной погрешности представляет собой трудоёмкую операцию, то на практике для обеспечения заданий точности все промежуточные вычисления проводят с одним или двумя запасными знаками.

1. Вычисление значений показательной функции.

Для показательной функции справедливо разложение

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} \quad (-\infty < x < \infty). \tag{2.4}$$

Вычисления удобно вести, пользуясь следующей рекуррентной записью:

$$e^x = \sum_{k=0}^{\infty} u_k \quad u_k = \frac{x}{k} u_{k-1}, \quad S_k = S_{k-1} + u_k \quad (k = 1,2,\ldots,n).$$

Где $u_0 = 1,\ S_0 = 1.$ Число $S_n = \sum_{k=0}^{n} \frac{x^k}{k!}$ приближённо даёт искомый результат e^x.

Для остатка ряда может быть получена следующая оценка :

$$|R_n(x)| < |u_n| \text{ при } 0 < 2|x| \le n;$$

поэтому процесс суммирования может быть прекращён, как только очередной вычисленный член ряда u_k будет по модулю меньше заданной допустимой погрешности ε :

$$|u_n| < \varepsilon \qquad \left(\textit{если только } |x| \le \frac{\pi}{2} \right).$$

При больших по модулю значениях x ряд $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} \quad (-\infty < x < \infty).$ (2.4) малопригоден для вычислений. В этих случаях обычно поступают так: представляют x в виде суммы

$$x = E(x) + q,$$

где $E(x) -$ целая часть x и $q -$ дробная его часть, $0 \le q < 1.$ Тогда

$$e^x = e^{E(x)} \cdot e^q.$$

Первый множитель $e^{E(x)}$ находится с помощью умножения

$$e^{E(x)} = \underbrace{e \cdot e \ldots e}_{E(x)\ раз}, \text{ если } E(x) > 0,$$

и

$$e^{E(x)} = \underbrace{\frac{1}{e} \cdot \frac{1}{e} \ldots \frac{1}{e}}_{-E(x)\ раз}, \text{ если } E(x) < 0,$$

Второй множитель вычисляется с помощью степенного разложения

$$e^q = \sum_{n=0}^{\infty} \frac{q^n}{n!}.$$

При $0 \le q < 1$ этот ряд быстро сходится, так как

$$0 \le R_n(q) < \frac{1}{n!n} q^{n+1}.$$

П р и м е р 2.1. Найти $\sqrt{e}$ с точностью до 10^{-5}.

Р е ш е н и е. Пользуемся формулой

$$e^{1/2} = \sum_{k=0}^{n} u_k + R_n\left(\frac{1}{2},\right) \tag{2.5}$$

где $u_0 = 1,\ u_k = \frac{u_{k-1}}{2k}$ $(k = 1,2,\ldots,n)$. Слагаемые подсчитываем с двумя запасными десятичными

знаками.
Последовательно имеем

$$u_0 = 1,$$

$$u_1 = \frac{u_0}{2} = 0,5000000,$$

$$u_2 = \frac{u_1}{4} = 0,1250000,$$

$$u_3 = \frac{u_2}{6} = 0,0208333,$$

$$u_4 = \frac{u_3}{8} = 0,0026042,$$

$$u_5 = \frac{u_4}{10} = 0,0002604,$$

$$u_6 = \frac{u_5}{12} = 0,0000217,$$

$$u_7 = \frac{u_6}{14} = \underline{0,0000016,}$$

$$S_7 = 1,6487212.$$

Округляя сумму до пяти десятичных знаков после запятой, получим

$$\sqrt{e} = 1,64872.$$

Для вычисления значений показательной функции $a^x \left(a > 0\right)$ можно использовать формулу $a^x = e^{x \ln a}$.

2. Вычисление значений синуса и косинуса.

Для вычисления значений функции $\sin x$ и $\cos x$ пользуемся степенными разложениями

$$\sin x = \sum_{k=0}^{\infty} \left(-1\right)^k \frac{x^{2k+1}}{\left(2k+1\right)!} \quad \left(-\infty < x < \infty\right), \tag{2.6}$$

$$\cos x = \sum_{k=0}^{\infty} \left(-1\right)^k \frac{x^{2k}}{\left(2k\right)!} \quad \left(-\infty < x < \infty\right), \tag{2.7}$$

Ряды (2.6) и (2.7) при больших x сходятся медленно, но, учитывая периодичность функций sin(x) и cos(x) и формулы приведения тригонометрических функций, легко заключить, что достаточно уметь вычислять sin(x) и cos(x) для промежутка $0 \le x \le \frac{\pi}{4}$. При этом можно использовать следующие рекуррентные формулы:

$$\left.\begin{aligned} \sin x &= \sum_{k=1}^{n} u_k + R_n\left(x\right), \\ u_1 = x, \;\; u_{k+1} &= -\frac{x^2}{2k\left(2k+1\right)} u_k \;\; \left(k = 1,2,\ldots,n-1\right); \end{aligned}\right\} \tag{2.8}$$

$$\left.\begin{aligned} \cos x &= \sum_{k=1}^{n} \upsilon_k + R_n(x), \\ \upsilon_1 = x,\ \ \upsilon_{k+1} &= -\frac{x^2}{2k(2k-1)} u_k \ \ (k = 1, 2, \ldots, n-1); \end{aligned}\right\} \qquad (2.9)$$

Так как в промежутке $\left(0, \frac{\pi}{4}\right)$ ряд (2.8) знакочередующийся с монотонно убывающими по модулю членами, то для его остатка R_n справедлива оценка

$$|R_n| \le \frac{|x|^{2n+1}}{(2n+1)!} = |u_{n+1}|.$$

Аналогично для ряда (2.5) $|R_n| \le |\upsilon_{n+1}|$. Следовательно, процесс вычисления $\sin x$ и $\cos x$ можно прекратить, как только очередной полученный член ряда по модулю будет меньше допустимой погрешности ε.

Пример 2.2. Вычислить $\sin 23^\circ 54'$ с точностью до 10^{-4}.

Решение. Переводим аргумент в радианы, сохраняя один запасной знак: $x = \text{arc}\, 23^\circ 54' = 0,41714.$ Применяя формулу, получим

$$u_1 = x = +0,41714,$$

$$u_2 = -\frac{x^2 u_1}{2 \cdot 3} = -0,01210,$$

$$u_3 = -\frac{x^2 u_2}{4 \cdot 5} = +0,00011,$$

$$u_4 = -\frac{x^2 u_3}{6 \cdot 7} = -0,00000.$$

Отсюда $\sin 23^\circ 54' = 0,40515 \approx 0,4052.$

Пример 2.3. Вычислить $\cos 17^\circ 24'$ с точностью 10^{-5}.

Решение. $x = \text{arc}\, 17^\circ 24' = 0,30369$. Применяя формулу (2.7), будем иметь

$$\upsilon_1 = 1,000000,$$

$$\upsilon_2 = -\frac{x^2}{1 \cdot 2} \upsilon_1 = -0,046114,$$

$$\upsilon_3 = -\frac{x^2}{3 \cdot 4} \upsilon_2 = +0,000354,$$

$$\upsilon_4 = -\frac{x^2}{5 \cdot 6} \upsilon_3 = -0,000001.$$

Отсюда

$$\cos 17^\circ 24' = 0,95424.$$

3. Вычисление значений гиперболического синуса и гиперболического косинуса.
Пользуемся степенными разложениями

$$\operatorname{sh} x = \sum_{k=1}^{\infty} \frac{x^{2k-1}}{(2k-1)!} \qquad (-\infty < x < \infty), \tag{2.10}$$

$$c\operatorname{h} x = \sum_{k=0}^{\infty} \frac{x^{2k}}{(2k)!} \qquad (-\infty < x < \infty) \tag{2.11}$$

и рекуррентной записью

$$\left.\begin{aligned} \operatorname{sh} x &= \sum_{k=1}^{\infty} u_k + R_n, \\ u_1 = x,\ u_{k+1} &= \frac{x^2}{2k(2k+1)} u_k; \end{aligned}\right\} \tag{2.12}$$

$$\left.\begin{aligned} c\operatorname{h} x &= \sum_{k=0}^{\infty} \upsilon_k + R_n^{\bullet}, \\ \upsilon_0 = 1,\ \upsilon_{k+1} &= \frac{x^2}{(2k+1)(2k+2)} \upsilon_k; \end{aligned}\right\} \tag{2.13}$$

При $n \geq |x| > 0$ имеют место оценки $R_n < \frac{1}{3}|u_n|$ и $R_n^* < \frac{2}{3}\upsilon$.

П р и м е р 2.4. Вычислить $\operatorname{sh} 1,4$ с точностью до 10^{-5}.
Р е ш е н и е. Применяя формулу (2.10), получим

$$u_1 = 1,400000,$$

$$u_2 = -\frac{x^2}{2 \cdot 3} u_1 = -0,4573333,$$

$$u_3 = -\frac{x^2}{4 \cdot 5} u_2 = +0,0448187,$$

$$u_4 = -\frac{x^2}{6 \cdot 7} u_3 = -0,0020915,$$

$$u_5 = -\frac{x^2}{8 \cdot 9} u_4 = 0,0000569,$$

$$u_6 = -\frac{x^2}{10 \cdot 11} u_5 = 0,0000010.$$

Отсюда

$$\operatorname{sh} 1,4 = 1,904301.$$

4. Вычисление значений логарифмической функции.
Пользуемся разложением по степеням $\frac{1-z}{1+z}$:

$$\ln z = -2\sum_{k=1}^{\infty} \frac{1}{2k-1} \left(\frac{1-z}{1+z}\right)^{2k-1} \qquad (0 < z < +\infty).$$

Пусть $x -$ положительное число. Представим его в виде

$$x = 2^m z,$$

где $m-$ целое число и $1/2 \le z < 1;$ тогда, полагая

$$\frac{1-z}{1+z} = \xi,$$

получим

$$\ln x = \ln 2^m z = m\ln 2 + \ln z = m\ln 2 - 2\sum_{k=1}^{\infty} \frac{1}{2k-1}\xi^{2k-1},$$

где $0 < \xi \le 1/2.$ Обозначив

$$u_k = \frac{\xi^{2k-1}}{2k-1} \qquad (k = 1,2,\ldots,n),$$

получаем рекуррентную запись

$$\left.\begin{aligned} \ln x &= m\ln 2 - 2\sum_{k=1}^{n} u_k + R_n, \\ u_1 &= \xi,\ u_{k+1} = \frac{(2k-1)\xi^2}{2k+1}u_k. \end{aligned}\right\} \tag{2.14}$$

Процесс суммирования прекращается, как только выполнится неравенство $u_n < 4\varepsilon,$ где $\varepsilon -$ допустимая погрешность.

ПРИМЕР 2.5. Найти ln5 с точностью до 10^{-6}.

Р е ш е н и е. Вычисления будем производить с двумя запасными знаками. Положим $5=2^3*0{,}625$. Следовательно, z=0,625 и $\xi = \frac{1-z}{1+z} = \frac{0{,}375}{1{,}625} = 0{,}23076923.$

Выпишем первые слагаемые:

$$\begin{aligned} u_1 &= \xi = 0{,}23076923, \\ u_2 &= \frac{\xi^3}{3} = 0{,}00409650, \\ u_3 &= \frac{\xi^5}{5} = 0{,}00013089, \\ u_4 &= \frac{\xi^7}{7} = 0{,}00000498, \\ \hline \textit{сумма} &= 0{,}23500160. \end{aligned}$$

По формуле (2.12) получим

$$\ln 5 = 3 \cdot 0{,}69314718 - 2 \cdot 0{,}23500160 = 1{,}609438.$$

З А Д А Ч И

1. Пользуясь разложением в степенной ряд, составить с указанной точностью ε таблицы значений следующих функций.

а) $e^x, \quad x = 0{,}300 + 0{,}002k \quad (k = 0,1,\ldots,14), \quad \varepsilon = 10^{-5},$

б) $e^x, \quad x = 2{,}500 + 0{,}002k \quad (k = 0,1,\ldots,14), \quad \varepsilon = 10^{-4},$

в) $e^{-x}, \quad x = 1{,}35 + 0{,}01k \quad (k = 0,1,\ldots,14), \quad \varepsilon = 10^{-5},$

г) $e^{-x}, \quad x = 0{,}505 + 0{,}005k \quad (k = 0,1,\ldots,15), \quad \varepsilon = 10^{-5},$

д) $e^{x^2}, \quad x = 0{,}50 + 0{,}02k \quad (k = 0,1,\ldots,15), \quad \varepsilon = 10^{-5},$

е) $e^{-x^2}, \quad x = 1{,}30 + 0{,}01k \quad (k = 0,1,\ldots,15), \quad \varepsilon = 10^{-5},$

ж) $\frac{1}{2\sqrt{n}}e^{\frac{-x^2}{2}}$, $x=1,78+0,03k$ $(k=0,1,\ldots,15)$, $x=2,545+0,005k$ $(k=0,1,\ldots,15)$.

2. Пользуясь разложением $\sin x$ и $\cos x$ в степенной ряд, составить таблицы значений следующих функций с точностью до 10^{-5}.

а) $\sin x$, $x=0,345+0,005k$ $(k=0,1,\ldots,15)$,

б) $\sin x$, $x=1,75+0,01k$ $(k=0,1,\ldots,15)$,

в) $\cos x$, $x=0,745+0,005k$ $(k=0,1,\ldots,15)$,

г) $\cos x$, $x=1,75+0,01k$ $(k=0,1,\ldots,15)$,

д) $\frac{\sin x}{x}$, $x=0,4+0,01k$ $(k=0,1,\ldots,15)$,

е) $\frac{\cos x}{x}$, $x=0,25+0,01k$ $(k=0,1,\ldots,15)$.

3. Пользуясь разложением $\operatorname{sh} x$ и $\operatorname{ch} x$ в степенной ряд, составить таблицы значений следующих функций с точностью до ε.

а) $\operatorname{sh} x$, $x=0,23+0,01k$ $(k=0,1,\ldots,15)$,

$\varepsilon=10^{-5}$, $x=23,0+0,05k$ $(k=0,1,\ldots,15)$, $\varepsilon=10^{-4}$,

б) $\operatorname{ch} x$ для тех же значений x.

§ 3. Некоторые многочленные приближения

Вычисление с помощью рядов Тейлора даёт достаточно быструю сходимость, вообще говоря, только при малых значениях $|x-x_0|$. Однако часто бывает нужно с помощью многочлена сравнительно невысокой степени подобрать приближение, которое давало бы достаточную точность для всех точек заданного отрезка. В этих случаях применяются разложения функций, полученные с помощью полиномов Чебышева на заданном отрезке. Ниже приводятся примеры таких разложений, указываются промежутки, в которых их следует использовать, а также соответствующие абсолютные погрешности ε. Для вычисления значений многочлена можно использовать схему Горнера.

1. Вычисление значений показательно функции на отрезке $[-1,1]$.

Пользуемся следующим многочленным приближением:

$$e^x \approx \sum_{k=0}^{7} a_k x^k \quad (|x|\le 1), \quad \varepsilon = 2\cdot 10^{-7}, \tag{2.15}$$

$a_0=0,9999998$, $a_1=1,0000000$, $a_2=0,5000063$, $a_3=0,1666674$,
$a_4=0,0416350$, $a_5=1,0083298$, $a_6=0,0014393$, $a_7=0,0002040$.

2. Вычисление значений логарифмической функции.

Имеет место формула

$$\ln(1+X)\approx\sum_{k=1}^{7}a_k x^k \quad (0\le x\le 1), \quad \varepsilon = 2{,}2\cdot 10^{-7}, \tag{2.16}$$

$a_1 = 0{,}999981028,\quad a_2 = -0{,}499470150\quad a_3 = 0{,}328233122,$
$a_4 = -0{,}225873284,\quad a_5 = -0{,}134639267,\quad a_6 = -0{,}055119959,$
$a_7 = 0{,}010757369.$

3. Вычисление значений тригонометрических функций.
Пользуемся следующими многочленными приближениями:

$$\sin x\approx\sum_{k=0}^{4}a_{2k+1}k^{2k+1} \quad \left(|x|\le\frac{\pi}{2}\right) \quad \varepsilon = 6\cdot 10^{-9}, \tag{2.17}$$

$a_1 = 1{,}000000002\quad a_3 = -0{,}166666589\quad a_5 = 0{,}008333075,$
$a_7 = -0{,}000198107,\quad a_9 = 0{,}000002608;$

$$\cos x\approx\sum_{k=0}^{5}a_{2k}x^{2k} \quad (|x|\le 1) \quad \varepsilon = 2\cdot 10^{-9}, \tag{2.18}$$

$a_0 = 1{,}000000000000,\quad a_2 = -0{,}499999999942,$
$a_4 = 0{,}041666665950,\quad a_6 = -0{,}001388885683,$
$a_8 = 0{,}000024795132,\quad a_{10} = -0{,}000000269591;$

$$\operatorname{tg} x\approx\sum_{k=0}^{6}a_{2k+1}x^{2k+1} \quad \left(|x|\le\frac{\pi}{4}\right) \quad \varepsilon = 2\cdot 10^{-8}, \tag{2.19}$$

$a_1 = 1{,}00000002,\quad a_3 = 0{,}33333082,\quad a_5 = 1{,}3339762,$
$a_7 = 0{,}05935836,\quad a_9 = 0{,}02457096,\quad a_{11} = 0{,}00294045,$
$a_{13} = 0{,}00947324.$

ПРИМЕР 2.6 Пользуясь многочленным приближением, найти значение $\sqrt{e}$ с точностью до 10^{-6}.
Р е ш е н и е. Вычисления проводим по формуле (2.15) пользуясь схемой Горнера при $x = 0{,}5$ (см. Таблица 2.2).

Таблица 2.2

$+\ 0{,}0002040$	$0{,}0014393$ $0{,}0001020$	$0{,}0083298$ $0{,}0007706$	$0{,}0416350$ $0{,}0045502$
$0{,}0002040$	$0{,}0015413$	$0{,}0091004$	$0{,}00461852$

$0{,}1666674$ $0{,}0230926$	$0{,}5000063$ $0{,}0948800$	$1{,}0000000$ $0{,}2974431$	$0{,}9999998\ \lfloor 0{,}5$ $0{,}6487216$
$0{,}1897600$	$0{,}5948863$	$1{,}2974431$	$\underline{1{,}6487214 = P(0{,}5)}$

Округляя до шести знаков после запятой, получим $e^{\frac{1}{2}} \approx 1{,}648721$ (ср. пример 2.1).

ПРИМЕР 2.7 Пользуясь многочленным приближением найти значение $\sin 0{,}5$ с точностью до 10^{-8}.
Р е ш е н и е. Вычисления можно провести по формуле (2.19), пользуясь схемой Горнера при $x = 0{,}5$. Но так как многочлен в формуле (2.19) содержит только нечётные степени x, удобнее переписать его в виде

$$\sum_{k=0}^{4} a_{2k+1}x^{2k+1} = \left(a_1 + a_3x^2 + a_5x^4 + a_7x^6 + a_9x^8\right)x$$

и применить схему Горнера к многочлену

$$P(\xi) = a_1 + a_3\xi + a_5\xi^2 + a_7\xi^3 + a_9\xi^4, \qquad \xi = x^2 = 0,25. \tag{2.20}$$

Вычисление значения $P(0,25)$ проведено в Таблица 2.3. Умножая полученное значение $P(0,25) = 0,958851087$ на $x = 0,5$ и округляя, получим искомое значение $\sin 0,5 \approx 0,47942554.$

Таблица 2.3

$0,000002608$	$-0,000198107$	$0,008333075$	$-0,166666589$	$0,000000002 \quad \underline{\vert 0,25}$
$+$	$0,000000652$	$-0,000049364$	$0,002070928$	$-0,041148915$
$0,000002608$	$-0,000197445$	$0,008283711$	$-0,164595661$	$\underline{0,958851087 = P(0,25)}$

З А Д А Ч И

1. Составить таблицы значений следующих функций с точностью до ε для указанных значений x.

а) $e^x, \quad x = 0,725 + 0,001k \quad (k = 0,1,2,\ldots,15), \quad \varepsilon = 10^{-4},$

$x = 0,213 + 0,002k \quad (k = 0,1,2,\ldots,15), \quad \varepsilon = 10^{-5},$

б) $e^x, \quad x = 0,213 + 0,003k \quad (k = 0,1,2,\ldots,15), \quad \varepsilon = 10^{-5},$

$x = 1,27 + 0,02k \quad (k = 0,1,2,\ldots,15), \quad \varepsilon = 10^{-5},$

в) $e^{\frac{1}{x}}, \quad x = 0,2 + 0,05k \quad (k = 0,1,2,\ldots,15), \quad \varepsilon = 10^{-5},$

г) $\dfrac{1}{2\sqrt{\pi}} e^{-\frac{x^2}{2}}, \quad x = 0,4 + 0,02k \quad (k = 0,1,2,\ldots,15), \quad \varepsilon = 10^{-5},$

$x = 1,2 + 0,1k \quad (k = 0,1,2,\ldots,15), \quad \varepsilon = 10^{-5}.$

2. Составить таблицы для следующих функций с точностью до 10^{-5} для указанных значений x.

а) $\sin x, \quad x = 0,055 + 0,003k \quad (k = 0,1,2,\ldots,15), \quad x = 0,80 + 0,05k \quad ,$
$(k = 0,1,2,\ldots,15),$

б) $\cos x$ для тех же значений x, в) $\operatorname{tg} x$ для тех же значений x.

§ 4. Применение цепных дробей для вычисления значений трансцендентных функций

1. **Вводные замечания**.

Пусть $a_0, a_1, a_2, \ldots, a_n, \ldots, b_1, b_2, \ldots, b_n, \ldots$ – две последовательности. Выражение вида

$$a_0 + \cfrac{b_1}{a_1 + \cfrac{b_2}{a_2 + \cfrac{b_3}{a_3 + \ddots}}} = \left[a_0; \frac{b_1}{a_1}, \frac{b_2}{a_2}, \frac{b_3}{a_3}, \ldots\right] \tag{2.21}$$

называется *цепной* или *непрерывной* дробью, отвечающей заданным последовательностям

$\{a_k\}_{k=0}^{\infty}$, $\{b_k\}_{k=1}^{\infty}$. В общем случае элементы цепной дроби a_0, a_k, b_k $(k=1,2,\ldots)-$ вещественные или комплексные числа или функции одной или нескольких переменных. Выражения

$$a_0+\frac{b_1}{a_1},\ a_0+\frac{b_1}{a_1+\dfrac{b_2}{a_2}}a_0,\ a_0+\frac{b_1}{a_1+\dfrac{b_2}{a_2+\dfrac{b_3}{a_3}}},\ldots,\ a_0+\cfrac{b_1}{a_1+\cfrac{b_2}{a_2+\ddots\ a_{n-1}+\cfrac{b_n}{a_n}}}$$

называются соответственно первой, второй, третьей, … , $n-$й *подходящей дробью* Для данной бесконечной дроби и обычно обозначаются $\frac{P_1}{Q_1}$, $\frac{P_2}{Q_2}$, … ,$\frac{P_n}{Q_n}$. При работе на ЭВМ подходящие цепные дроби удобно находить с помощью следующей последовательности операций:

$$c_1=\frac{b_n}{a_n},\qquad d_1=a_{n-1}+c_1,$$
$$c_2=\frac{b_{c-1}}{d_1},\quad d_2=a_{n-2}+c_2,$$
$$\ldots\ldots\ldots\ldots\ldots$$
$$c_k=\frac{b_{n-k+1}}{d_{k-1}},\quad d_k=a_{n-k}+c_k,$$
$$\ldots\ldots\ldots\ldots\ldots$$
$$c_n=\frac{b_1}{d_{n-1}},\quad d_n=a_0+c_n=\frac{P_n}{Q_n}.$$

Указанная последовательность действий легко программируется.
Для числителей и знаменателей подходящих дробей имеют место следующие рекуррентные формулы:

$$P_n=a_nP_{n-1}+b_nP_{n-2},\qquad Q_n=a_nQ_{n-1}+b_nQ_{n-2},\qquad (n=1,2,\ldots),$$

где $P_{-1}=1,\ Q_{-1}=0,\ P_0=a_0,\ Q_0=1.$

Цепная дробь (2.21) называется *сходящейся*, если существует конечный предел A подходящей дроби P_n/Q_n при $n\to\infty$:

$$A=\lim_{n\to\infty}\frac{P_n}{Q_n},$$

причём число A принимается значение этой дроби. Если же предел не существует, то цепная дробь (2.21) называется *расходящейся*. Для сходящейся цепной дроби подходящая дробь P_n/Q_n является её приближённым значением.

2. Разложение $\mathbf{e^x}$ в цепную дробь. Пользуемся следующим разложением:

$$e^x=\left[0,\ \frac{1}{1},\ \frac{-2x}{2+x},\ \frac{x^2}{6},\ \frac{x^2}{10},\ldots,\ \frac{x^2}{4n+2},\ldots\right]. \qquad (2.22)$$

Для любого x эта дробь сходящаяся. Подходящими дробями для данной функции являются

$$\frac{P_1}{Q_1}=\frac{1}{1},\qquad \frac{P_2}{Q_2}=\frac{2+x}{2-x},\qquad \frac{P_3}{Q_3}=\frac{12+6x+x^2}{12-6x+x^2},$$

$$\frac{P_4}{Q_4}=\frac{120+60x+12x^2+x^3}{120-60x+12x^2-x^3},$$

$$\frac{P_5}{Q_5}=\frac{1680+840x+180x^2+20x^3+x^4}{1680-840x+180x^2-20x^3+x^4},$$

П р и м е р 2.8. Пользуясь разложением e^x в цепную дробь, вычислить $1/e$ с точностью до 10^{-5}.

Р е ш е н и е. Вычислим четвёртую и пятую подходящие дроби при $x=-1$:

$$\frac{P_4}{Q_4}=\frac{120-60+12-1}{120+60+12+1}=\frac{71}{193}\approx 0,367876.$$

$$\frac{P_5}{Q_5}=\frac{1680-840+180-20+1}{1680+840+180+20+1}=\frac{1001}{2721}\approx 0,367879.$$

Сравнивая полученные приближения, замечаем, что у них совпадают пять десятичных знаков, поэтому можно положить $e^{-1}\approx 0,36788.$

3. Разложение $\mathbf{tgx}$ в цепную дробь.

Имеет место следующее разложение:

$$\operatorname{tg}x=\left[0,\ \frac{x}{1},\ -\frac{x^2}{3},\ -\frac{x^2}{5},\ldots,\ 0\frac{x^2}{2n+1},\ldots\right]. \tag{2.23}$$

Это разложение справедливо во всех точках непрерывности $\operatorname{tg}x$. Первыми подходящими дробями будут

$$\frac{P_1}{Q_1}=\frac{1}{1},\qquad \frac{P_4}{Q_4}=\frac{105x-10x^3}{105-45x^2+x^4},$$

$$\frac{P_2}{Q_2}=\frac{3x}{3-x^2},\qquad \frac{P_5}{Q_5}=\frac{945x-105x^3+x^5}{945-420x^2+15x^4},$$

$$\frac{P_3}{Q_3}=\frac{15x-x^3}{15-6x^2},$$

З А Д А Ч И

1. Пользуясь разложениями в цепную дробь, составить таблицы значений следующих функций с точностью до ε.

а) $e^x,\quad x=0,155+0,005k\quad (k=0,1,2,\ldots,15),\quad \varepsilon=10^{-5},\quad x=0,30+0,03k$

$(k=0,1,2,\ldots,15),\ \varepsilon=10^{-4},$

б) $\operatorname{tg}x,\quad x=0,47+0,005k\quad (k=0,1,2,\ldots,15),\quad \varepsilon=10^{-5}$

§ 5. Применение метода итераций для приближённого вычисления значений функций

Всякую функцию $y=f(x)$ можно различными способами задавать неявно, т.е. некоторым уравнением

$$F(x,y)=0. \tag{2.24}$$

Часто бывает, что решение уравнения (2.24) относительно y каким-либо итерационным методом сводится к однообразным операциям, легко реализуемым на ЭВМ. Тогда, очевидно, целесообразно применить метод итераций.

Один из возможных итерационных процессов для вычисления $y(x)$ можно построить следующим образом.

Пусть y_n – приближённое значение y. Применив формулу Лагранжа, получим

$$F(x,y_n)=F(x,y_n)-F(x,y)=(y_n-y)F_y'(x,\overline{y}_n),$$

где $\overline{y}_n$ – некоторое промежуточное значение между y_n и y. Отсюда

$$y=y_n-\frac{F(x,y_n)}{F_y'(x,\overline{y}_n)},$$

Причём значение $\overline{y}_n$ нам не известно.

Полагая приближённо $\overline{y}_n\approx y_n$, получим следующую формулу для вычисления $y\approx y_{n+1}$:

$$y_{n+1}=y_n-\frac{F(x,y_n)}{F_y'(x,y_n)} \qquad (n=0,1,2,\ldots). \tag{2.25}$$

Если $F_y'(x,y)$ и $F_{yy}''(x,y)$ существуют и сохраняют постоянные знаки в рассматриваемом интервале, содержащем корень $y(x)$, то итерационный процесс сходится к $y(x)$.

Начальное приближение $y_0(x)$ выбирают так, чтобы оно легко вычислялось и было, по возможности, близким к истинному значению $y(x)$.

Процесс итераций продолжается до тех пор, пока в пределах заданной точности два последовательных значения y_{n+1} и y_n не совпадут между собой, после чего приближённо полагают

$$y(x)\approx y_{n+1}.$$

1. Вычисление обратной величины. Пусть $y=1/x\ (x>0)$. Положим $F(x,y)=x-1/y=0.$ Применив формулу (2.25), получим

$$y_{n+1}=y_n(2-xy_n). \tag{2.26}$$

Вычисление y_{n+1} по полученной итерационной формуле (2.26) содержит лишь действия умножения и вычитания. Таким образом, можно находить 1/x на вычислительных машинах, в которых нет операции деления. Начальное значение y_0 выбирается обычно следующим образом. Записывают аргумент x в двоичной системе:

$$x=2^m x_1,$$

где m – целое число и $1/2\le x_1<1$. Полагают $y_0=2^{-m};$ при таком выборе начального значения y_0 сходимость итерационного процесс довольно быстрая.

Замечание. Так как частное a/b есть произведение a на $1/b$, то деление на машинах, в которых нет операции деления, можно реализовать в два этапа:

1) Вычисление $y=1/b$ (обратной величины делителя),

2) Умножение y на делимое a.

ПРИМЕР 2.9. С помощью формулы (2.26) найти значение функции $y=1/x$ при $x=5$ с точностью до 10^{-4}.

Решение. Запишем аргумент x в виде $x=2^3\cdot\frac{5}{8}$. Полагаем $y_0=2^{-3}=1/8$. По формуле (2.23) будем иметь

$$y_1=\frac{1}{8}\left(2-\frac{5}{8}\right)=\frac{11}{64}=0{,}1718, \qquad y_2=\frac{11}{64}\left(2-\frac{55}{64}\right)=\frac{803}{4096}=0{,}1960,$$

$$y_3 = 0{,}1960\left(2 - 0{,}9800\right) = 0{,}1960 \cdot 1{,}0200 = 0{,}19992.$$

Мы видим, что здесь уже третье приближение даёт $y\left(x\right) \approx 0{,}1999$ с точностью до 10^{-4}.

2. Вычисление квадратного корня. Пусть $y = \sqrt{x}\ \ \left(x > 0\right).$ Преобразуем это уравнение к виду

$$F\left(x, y\right) \equiv y^2 - x = 0.$$

Применяя формулу (2.22), получим

$$y_{n+1} = \frac{1}{2}\left(y_n + \frac{x}{y_n}\right) \qquad \left(n = 0,1,2,\ldots\right). \tag{2.27}$$

Эта формула называется *формулой Герона.*

Пусть аргумент x записан в двоичной системе:

$$x = 2^m x_1,$$

где $m -$ целое число и $1/2 \le x_1 < 1.$ Тогда обычно полагают

$$y_0 = 2^{E\left(m/2\right)},$$

Где $E\left(m/2\right) -$ целая часть числа $m/2$.

Итерационный процесс по формуле Герона легко реализуется на машине, имеющей деление в качестве элементарной операции; при этом процесс итераций сходится при любом выборе $y_0 > 0$ (в этом примере легко проверить выполнение указанных выше условий сходимости, так как $F_y' = 2y > 0$ и $F_{yy}'' = 2y > 0$).

Если $0{,}01 \le x \le 1,$ то за начальное приближение можно брать $y_0 = ax + b;$ соответствующие коэффициенты a и b приведены в Таблица 2.4.

Таблица 2.4

Коэффициенты для начального приближения в формуле Герона

Интервал	a	b	Интервал	a	b
$\left(0{,}01;0{,}02\right)$	4,1	0,060	$\left(0{,}18;0{,}30\right)$	1,0	0,247
$\left(0{,}02;0{,}03\right)$	3,2	0,078	$\left(0{,}30;0{,}60\right)$	0,8	0,304
$\left(0{,}03;0{,}08\right)$	2,2	0,110	$\left(0{,}60;1{,}00\right)$	0,6	0,409
$\left(0{,}08;0{,}018\right)$	1,4	0,174			

При таком выборе начального приближения y_0 уже вторая итерация y_2 даёт значение $\sqrt{x}$ с восемью десятичными знаками после запятой, причём при вычислении y_0 можно брать значение x лишь с тремя десятичными знаками.

ПРИМЕР 2.10. Найти $\sqrt{7}$ с точностью до 10^{-5}.

Р е ш е н и е. Здесь $x = 7 = 2^3 \cdot \frac{7}{8}$. Следовательно, начальное приближение имеет вид

$$y_0 = 2^{E\left(3/2\right)} = 2.$$

По формуле (2.24) последовательно находим

$$y_1 = \frac{1}{2}\left(2 + \frac{7}{2}\right) = 2,75000,$$

$$y_2 = \frac{1}{2}\left(\frac{11}{4} + \frac{28}{11}\right) = 2,64772,$$

$$y_3 = \frac{1}{2}\left(\frac{233}{88} + \frac{616}{233}\right) = 2,64575,$$

$$y_4 = \frac{1}{2}\left(2,64575 - \frac{7}{2,64575}\right) = 2,64575.$$

Заметим, что в значениях y_3 и y_4 совпадают пять десятичных знаков, приближённо полагаем $\sqrt{7} \approx 2,64575$.

Замечание. Если вычисление ведётся на вычислительной машине, система команд которой не содержит операции деления, то можно пользоваться другой итерационной формулой, а именно:

$$y_{n+1} = y_n\left(\frac{3}{2} - \frac{y_n^2}{2x}\right) \qquad (n = 0,1,2,\ldots). \tag{2.28}$$

Извлечение квадратного корня сводится по этой формуле к однократному вычислению обратной величины $\frac{1}{2x}$ и затем к итерационному процессу, каждый этап которого содержит лишь действия умножения и вычитания. Формула (2.28) соответствует преобразованию исходного уравнения к виду

$$F(x,y) \equiv \frac{1}{y^2} - \frac{1}{x} = 0.$$

3. Вычисление обратной величины квадратного корня. Пусть имеем

$$y = \frac{1}{\sqrt{x}} \qquad (x > 0).$$

Итерационная формула для вычисления обратной величины квадратного корня имеет вид

$$y_{n+1} = \frac{3}{2}y_n - \frac{1}{2}xy_n^3 \qquad (n = 0,1,2\ldots). \tag{2.29}$$

Формула (2.29) получается при преобразовании исходного уравнения $y = \frac{1}{\sqrt{x}}$ к виду

$$F(x,y) \equiv \frac{1}{y^2} - x = 0.$$

В качестве начального приближения обычно берут

$$y_0 = 2^{-E(m/2)},$$

где $x = 2^m x_1,\ 1/2 \le x_1 < 1$.

Мы имеем здесь итеративный процесс также «без деления».

4. Вычисление кубического корня. Пусть имеем $y = \sqrt[3]{x}$. Применив формулу (2.25) к уравнению $F(x,y) \equiv y^3 - x = 0$, получим итерационную формулу для вычисления кубического корня в виде

$$y_{n+1}=\frac{1}{3}\left(\frac{2y_n^2+x}{y_n^2}\right); \tag{2.30}$$

начальное приближение

$$y_0=2^{E(m/2)},$$

где $x=2^m x_1,\ m-$ целое число и $1/2\le x_1<1.$

ПРИМЕР 2.11. Вычислить $\sqrt[3]{5}$ с точностью до 10^{-3}.

Р е ш е н и е. Здесь $x=5=2^3\cdot\frac{5}{8}$. Начальное приближение

$$y_0=2^{E(3/3)}=2,$$

$$y_1=\frac{1}{3}\left(4+\frac{5}{4}\right)=\frac{21}{12}=1{,}7500.$$

Дальнейшие вычисления сведены в Таблица 2.5.

Таблица 2.5 Вычисление $\sqrt[3]{5}$

n	y_n	y_n^2	$3y_n^2$	y_n^3	$2y_n^3+5$
0	2	4	12	8	21
1	1,7500	3,0625	9,1875	5,3594	15,7188
2	1,7100	2,9241	8,7723	5,0002	15,0004
3	1,7100				

Таким образом, $\sqrt[3]{5}\approx 1{,}710.$

5. Вычисление корня $p-$й степени. Пусть

$$y=\sqrt[p]{x},$$

где $x>0$ и $p>0-$ целое число.

Применив формулу (2.22) и уравнению $F(x,y)\equiv 1-\frac{x}{y^F}=0,$ получим

$$y_{n+1}=y_n\left[\left(1+\frac{1}{p}\right)-\frac{y_n^p}{px}\right]. \tag{2.31}$$

Итерационный процесс будет сходящимся, если только начальное приближение $y_0>0$ выбрать настолько малым, чтобы $y_0^p<(p+1)x.$

6. **Формула Ньютона для вычисления корня p-й степени.** Пусть $y=\sqrt[p]{x};$ тогда имеет место формула

$$y_{n+1}=\frac{1}{p}\left[(p-1)y_n+\frac{x}{y_n^{p-1}}\right], \tag{2.32}$$

получающаяся из формулы (2.28) при $F(x,y)\equiv y^p-x.$

Начальное приближение y_0 можно подобрать с точностью до одной-двух значащих цифр.

При p=2 из формулы Ньютона получаем формулу Герона.

ПРИМЕР 2.12. Вычислить $y=\sqrt[7]{277\,234}$ с точностью до 10^{-6}.

Р е ш е н и е. Возьмём $y_0=6;$ по формуле (5.8) последовательно вычисляем:

$$y_1=6\left[\left(1+\frac{1}{7}\right)-\frac{6^7}{7\cdot 277\,234}\right]=5,99164605,$$

$$y_2=5,99169225,\quad y_3=5,99169225.$$

Ответ: $\sqrt[7]{277\,234}\approx 5,991692.$

ПРИМЕР 2.13. Вычислить $y=\sqrt[4,78]{16234}$ с точностью до 10^{-8}.

Решение. Возьмём $y_0=7;$ тогда по формуле (2.28) будем иметь

$$y_1=\frac{1}{4,78}\left[(4,78-1)\cdot 7+\frac{16234}{7\cdot 3,78}\right]=7,70590133,$$

$$y_2=7,60319046,\quad y_3=7,60050180,\quad y_4=7,60050001,\quad y_5=7,60050001.$$

Таким образом, с точностью до 10^{-6} получаем

$$\sqrt[4,78]{16\,234}\approx 7,600500.$$

З А Д А Ч И

1. Пользуясь методом итераций, составить таблицы значений следующих функций с точностью до 10^{-6}.

а) $1/x,\quad x=3+2k\quad (k=0,1,2,\ldots,15)$, б) $1/x^2$ для тех же значений x, в) $1/x^3$ для тех же значений x, г) $\dfrac{x}{1+x}$, $x=0,007+0,003k\ (k=0,1,2,\ldots 15).$

2. Пользуясь методом итераций, составить таблицы значений следующих функций с точностью до 10^{-5}.

а) $\sqrt{x},\quad x=2+k\quad (k=0,1,2,\ldots,15),$

б) $x\sqrt{x}$ для тех же значений x,

в) $\sqrt{1+x^2},\quad x=0,3+0,002k\quad (k=0,1,2,\ldots,15),$

г) $\sqrt{x^2+\dfrac{1}{x}}$ для тех же значений x,

3. Пользуясь методом итераций поставить таблицы значений следующих функций с точностью до 10^{-5}.

а) $1/\sqrt{x},\ x-3+3k\quad (k=0,1,2,\ldots,15),$

б) $1/\sqrt{2+x^2},\quad x=0,3+0,002k\quad (k=0,1,2,\ldots,15),$

в) $(2x+1)/\sqrt{x^2},\quad x=3,1+0,005k\quad (k=0,1,2,\ldots,15),$

г) $1/\sqrt{x(x+1)},\ x=2,3+0,002k \quad (k=0,1,2,\ldots,15),$

4. Пользуясь методом итераций, составить таблицы значений функций с точностью до 10^{-6}.

а) $\sqrt[3]{x}, \quad x=3+k \quad (k=0,1,2,\ldots 15),$

б) $1/\sqrt[3]{x}$ для тех же значений x.

5. Пользуясь методом итераций, составить таблицы значений следующих функций с точностью до 10^{-6}.

а) $\sqrt[4]{x}, \quad x=0,05+0,02k \quad (k=0,1,2,\ldots,15),$

б) $\sqrt[5]{x}$ для тех же значений x,

в) $\sqrt[6]{x}$ для тех же значений x,

г) $\sqrt[7]{x}$ для тех же значений x.

Глава 3
РЕШЕНИЕ НЕЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

Пусть f- полином или трансцендентная функция одного переменного, действительного или комплексного. Задача состоит в том, чтобы найти один или более *нулей* f, т. е. решений уравнения f(x)=0. Нахождение формул для нулей полиномов было одним из важнейших разделов итальянской математики эпохи Ренессанса. Для полиномов 2-й, 3-й и 4-й степеней ещё несколько столетий назад были найдены алгоритмы, выражающие корни посредством конечного числа квадратичных или кубичных радикалов и рациональных операций. Но только в тридцатых года прошлого века Галуа доказал невозможность подобных алгоритмов для полиномов 5-й или более высокой степени, даже если допустить в формулах радикалы с показателем n.

§ 1. Уравнения с одним неизвестным. Метод деления пополам. Метод хорд. Метод касательной. Метод простой итерации.

1.1 Уравнения с одним неизвестным

Вводные замечания

Нахождения корней нелинейных уравнений вида

F(x)=0

1.1.1 Метод деления пополам

Пусть дано уравнение

$$F(x)=0 \qquad (3.1)$$

где функция *f(x)* непрерывна на *[a, b]* и *f(a)f(b)* < 0. Для нахождения корня уравнения (3.1), принадлежащего отрезку *[a, b]*, делим этот отрезок пополам. Если $f(\frac{a+b}{2})=0$, т $\xi=\frac{a+b}{2}$ является корнем уравнения. Если $f=\frac{a+b}{2}\neq 0$, то выбираем ту из половин $\left[a,\frac{a+b}{2}\right]$ или $\left[\frac{a+b}{2},b\right]$, на концах которой функция *f(x)* имеет противоположные знаки. Новый суженный отрезок $[a_1,b_1]$ снова делим пополам и проводим то же рассмотрение и т. д. В результате получаем на каком-то этапе или точный корень уравнения (3.1), или же бесконечную последовательность вложенных друг в друга отрезков $[a_1,b_1]$, $[a_2,b_2]$, . . . , $[a_n,b_n]$, . . . таких, что

$$f(a_n)\,f(b_n)<0\,(n=1,2,\ldots) \qquad (3.2)$$

$$b_n-a_n=\frac{1}{2^n}(b-a). \qquad (3.3)$$

Так как левые концы $a_1,a_2,\ldots a_n,\ldots$ образуют монотонную неубывающую ограниченную последовательность, а правые концы $b_1,b_2,\ldots b_n,\ldots$ —монотонную невозрастающую ограниченную последовательность, то в силу равенства (

(3.3) существует общий предел

$$\xi=\lim_{n\to\infty}a_n=\lim_{n\to\infty}b_n$$

Переходя к пределу при $n\to\infty$ в неравенстве (3.2), в силу непрерывности функции *f(x)* получим $[f(\xi)]^2\le 0$. Отсюда $f(\xi)=0$, т. е. ξ является корнем уравнения , причем, очевидно,

$$0\le\xi-a_n\le\frac{1}{2^n}(b-a) \qquad (3.4)$$

Если корни уравнения (3.1) не отделены на отрезке [a, b], то таким способом можно найти один из корней уравнения (3.3).

Метод половинного деления практически удобно применять для грубого нахождения корня данного уравнения, так как при увеличении точности значительно возрастает объем вычислительной работы.

Заметим, что метод половинного деления легко реализуется на электронных счетных машинах. Программа вычисления составляется так, чтобы машина находила значение правой части уравнения (3.3) в середине каждого из отрезков $[a_n,b_n]\ (n=1,2,\ldots)$ и выбирала соответствующую половину его.

Пример. Методом половинного деления уточнить корень уравнения

$$f\left(x\right)\equiv x^4+x^2-x-1=0,$$

лежащий на отрезке [0,1].

Решение. Последовательно имеем:

f(0)=-1; f(1)=1;
f(0,5)=0,06+0,25-0,5-1=-1,19;
f(0,75)=0,32+0,84-0,75-1=-0,59;
f(0,875)=0,59+1,072-0,812-1=-0,304;
f(0,8438)=0,507+1,202-0,844-1=-0,135;
f(0,8594)=0,546+1,270-0,859-1=-0,043 и т.д.

Можно принять

$$\xi=\frac{1}{2}\left(0,859+0,875\right)=0,867.$$

1.1.2 Метод хорд.

Укажем более быстрый способ нахождения корня ξ уравнения *f(x) = 0,* лежащего на заданном отрезке [a, b] таком, что *f(a)f(b)* < 0.

Пусть для определенности *f(a)*< 0 и *f(b)*> 0. Тогда, вместо того чтобы делить отрезок *[a,b]* пополам, более естественно разделить его в отношении —*f(a):f(b).* Это дает нам приближённое значение корня

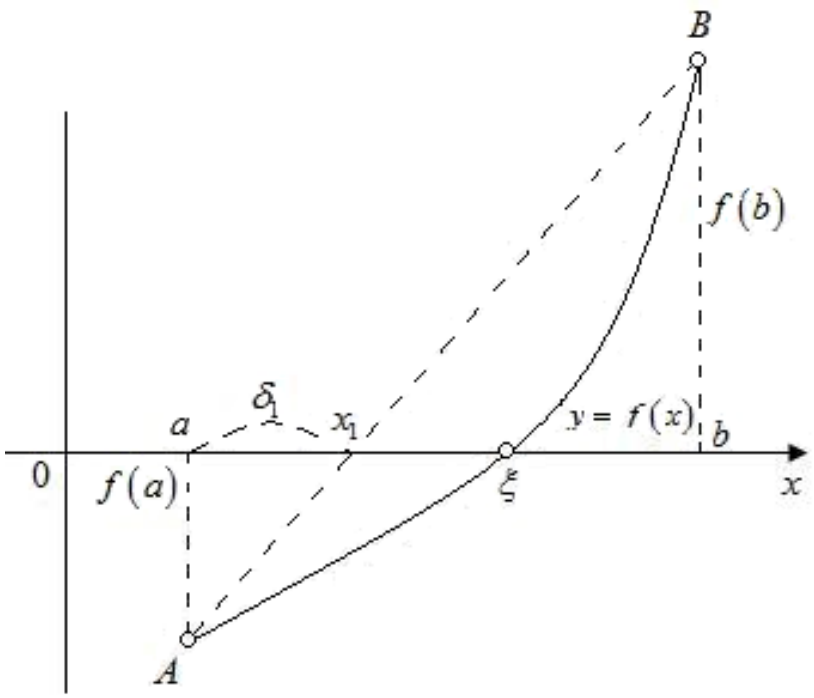

Рисунок 3.1

$$x_1=a+h_1, \tag{3.5}$$

где

$$h_1=\frac{-f\left(a\right)}{-f\left(a\right)+f\left(b\right)}\left(b-a\right)=-\frac{f(a)}{f(b)-f(a)}\left(b-a\right) \tag{3.6}$$

Далее, применяя этот прием к тому из отрезков [a,x_1] или [x_1, b], на концах которого функция *f*(x) имеет противоположные знаки**,** получим второе приближение корня x_2 и т. д.

Геометрически способ пропорциональных частей эквивалентен замене кривой *y=f(x)* хордой, проходящей через точки (**Рисунок 3.1**). В самом деле, уравнение хорды *AB* есть

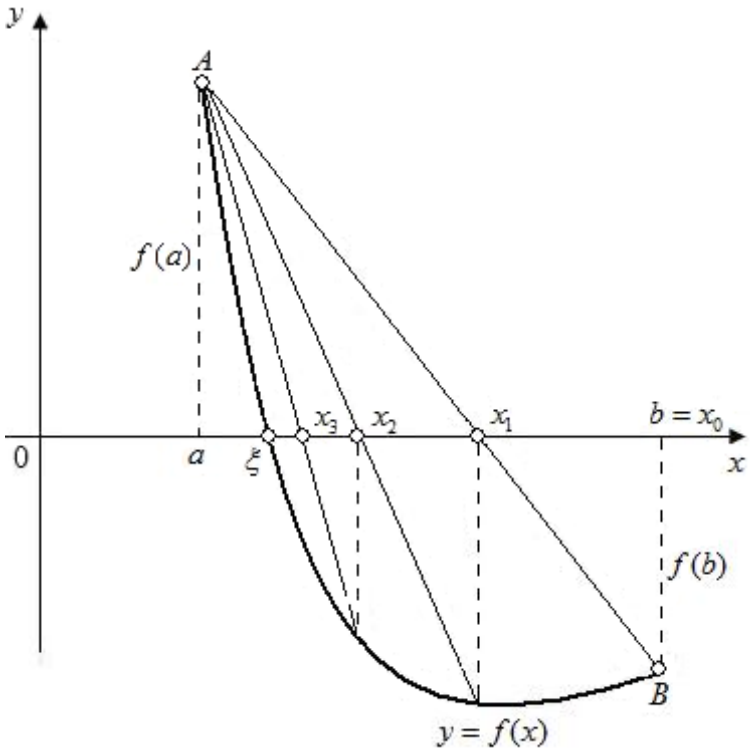

Рисунок 3.2

$$\frac{x-a}{b-a}=\frac{y-f(a)}{\gamma(b)-\gamma(a)}$$

Отсюда, полагая x = x_1 и y=0 , получим:

$$x_1=a-\frac{f(a)}{\gamma(b)-\gamma(a)}(b-a). \tag{3.7}$$

Формула (3.7) полностью эквивалентна формулам (3.3) и (3.4). Для доказательства сходимости процесса предположим,что корень отделен и вторая производная *f''(x)* сохраняет постоянный знак на отрезке [a, b]. Пусть для определенности *f''(x)> 0* при *a≤x≤b* (случай *f''(x)<0* сводится к нашему, если записать уравнение в виде *– f(x)=0*).

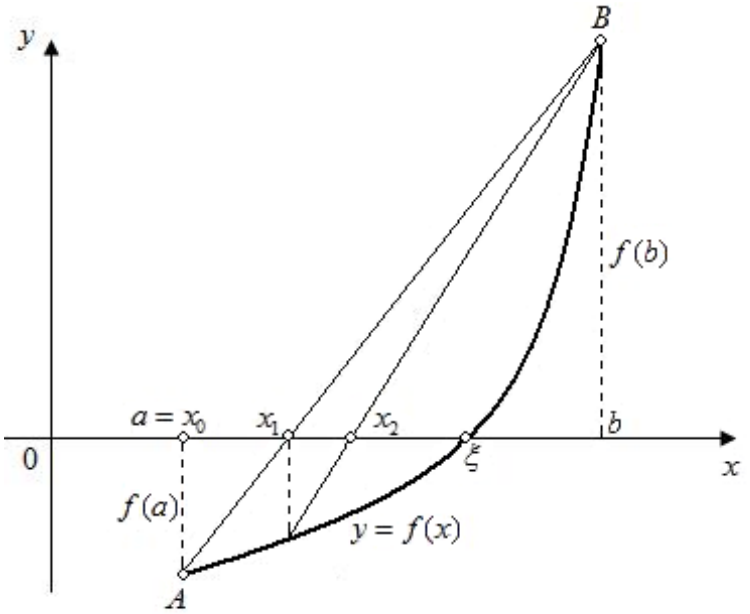

Рисунок 3.3

Тогда кривая *y=f(x)* будет выпукла вниз и, следовательно, расположена ниже своей хорды AB. Возможны два случая: 1) *f*(a)>0 (рис. 3.3) и 2) *f*(a)<0 (рис. 3.2).

В первом случае конец *a* неподвижен и последовательные приближения: $x_0 = b$;

$$x_{n+1} = x_n - \frac{f(x_n)}{\gamma(x_n) - \gamma(a)}(x_n - a)x \qquad (n = 0,1,2\ldots) \qquad (3.8)$$

образуют ограниченную монотонно убывающую последовательность, причем

$$a < \xi < \ldots < x_{n+1} < x_n < \ldots < x_1 < x_0$$

Во втором случае неподвижен конец *b*, а последовательные приближения: $x_0 = a$;

$$x_{n+1} = x_n - \frac{f(x_n)}{\gamma(b) - \gamma(x_n)}(b - x_n) \qquad (3.9)$$

образуют ограниченную монотонно возрастающую последовательность, причем

$$x_0 < x_1 < x_2 < \ldots < x_n < x_{n+1} < \ldots < \xi < b.$$

Обобщая эти результаты, заключаем: 1) неподвижен тот конец, для которого знак функции *f*(x) совпадает со знаком ее второй производной *f*''(x); 2) последовательные приближения x_n лежат по ту сторону корня ξ, где функция *f*(x) имеет знак, противоположный знаку ее второй производной *f*''(x). В обоих случаях каждое следующее приближение x_{n+1} ближе к корню ξ, чем предшествующее x_n. Пусть

$$\overline{\xi} = \lim_{n\to\infty} x_n \qquad \left(a < \overline{\xi} < b\right)$$

(предел существует, так как последовательность $\{x_n\}$ ограничена и монотонна). Переходя к пределу в равенстве (3.8), для первого случая будем иметь:

$$\overline{\xi} = \overline{\xi} - \frac{f(\overline{\xi})}{\gamma(\overline{\xi}) - \gamma(a)}(\overline{\xi} - a);$$

Отсюда $f(\overline{\xi}) = 0.$ Так как по предположению уравнение *f*(x)=0 имеет единственный корень ξ на интервале (*a, b*), то, следовательно, $\overline{\xi} = \xi$, что и требовалось доказать.

Совершенно так же переходом к пределу в равенстве (3.9) доказывается, что $\overline{\xi} = \xi$ для второго случая.

Для оценки точности приближения можно воспользоваться формулой:

$$|x_n - \xi| \le \frac{|f(x_n)|}{m_1}$$

где $|f'(x)| \ge m_1$ при $a \le x \le b$.

Приведем еще формулу, позволяющую оценивать абсолютную погрешность приближенного значения x_n, если известны два последовательных приближения x_{n-1} и x_n.

Будем предполагать, что производная *f*'(x) непрерывна на отрезке [*a, b*], содержащем все приближения, и сохраняет постоянный знак, причем

$$0 < m_1 \le |f'(x)| \le M_1 < +\infty. \qquad (3.10)$$

Примем для определенности, что последовательные приближения x_n точного корня ξ вырабатываются по формуле (3.8) (рассмотрение формулы (3.9) аналогично)

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{\gamma(x_n - 1) - \gamma(a)}(x_{n-1} - a)$$

(*n*= 1, 2, ...), где конец *a* является неподвижным. Отсюда, учитывая, что $f(\xi) = 0$ будем иметь:

$$f(\xi) - f(x_{n-1}) = \frac{f(x_{n-1}) - f(a)}{x_{n-1} - a}(x_n - x_{n-1}).$$

Применяя теорему Лагранжа о конечном приращении функции, получим:

$$(\xi - x_{n-1}) f'(\xi_{n-1}) = (x_n - x_{n-1}) f'(\overline{x}_{n-1}),$$

где $\xi_{n-1} \in (x_{n-1}, \xi)$ и $\overline{x}_{n-1} \in (a, x_{n-1}).$

Следовательно,

$$\left|\xi - x_n\right| = \frac{\left|f'(\overline{x}_{n-1}) - f'(\xi_{n-1})\right|}{\left|\gamma'(\xi_{n-1})\right|}\left|x_n - x_{n-1}\right|. \qquad (3.11)$$

Так как $f'(x)$ сохраняет постоянный знак на отрезке $[a, b]$, причем $\overline{x}_{n-1} \in [a,b]$ и $\xi_{n-1} \in [a,b]$, очевидно, имеем:

$$\left|f'(\overline{x}_{n-1}) - f'(\xi_{n-1})\right| \le M_1 - m_1.$$

Поэтому из формулы (3.11) выводим:

$$\left|\xi - x_n\right| \le \frac{M_1 - m_1}{m_1}\left|x_n - x_{n-1}\right|, \qquad (3.12)$$

где за m_1 и M_1 могут быть взяты соответственно наименьшее и наибольшее значения модуля производной $f'(x)$ на отрезке $[a, b]$. Если отрезок $[a, b]$ столь узок, что имеет место неравенство

$$M_1 \le 2m_1,$$

то из формулы (3.7) получаем:

$$\left|\xi - x_n\right| \le \left|x_n - x_{n-1}\right|.$$

Таким образом, в этом случае, как только будет обнаружено, что

$$\left|x_n - x_{n-1}\right| < \varepsilon,$$

где ξ — заданная предельная абсолютная погрешность, то гарантировано, что

$$\left|\xi - x_n\right| < \varepsilon.$$

Пример. Найти положительный корень уравнения

$$f(x) \equiv x^3 - 0,2x^3 - 0,2x - 1,2 = 0$$

с точностью до 0,002.

Решение. Прежде всего отделяем корень. Так как

$f(1)=-0,6<0$ и $f(2)=5,6>0$,

то искомый корень ξ лежит в интервале (1, 2). Полученный интервал велик, поэтому разделим его пополам. Так как

$$f(1,5) = 1,425, \qquad 1 < \xi < 1,5.$$

Последовательно применяя формулы (1) и (2), будем иметь:

$$x_1 = 1 + \frac{0,6}{1,425 + 0,6}(1,5 - 1) = 1 + 0,15 = 1,15;$$

$$f(x_1) = -0,173;$$

$$x_2 = 1,15 + \frac{0,173}{1,425 + 0,073}(1,5 - 1,15) = 1,15 + 0,040 = 1,190;$$

$$f(x_2) = -0,036;$$

$$x_3 = 1,190 + \frac{0,036}{1,425 + 0,036}(1,5 - 1,190) = 1,190 + 0,008 = 1,198;$$

$$f(x_3) = -0,0072.$$

Так как $f'(x) = 3x^2 - 0,4x - 0,2$ и при $x_3<x<1,5$ имеем

$$f'(x) = 3x^2 - 0,4x - 0,2 = 3 \cdot 1,43 - 0,8 = 3,49,$$

то можно принять:

$$0 < \xi - x_3 < \frac{0,0072}{3,49} \approx 0,002.$$

Таким образом, $\xi = 1,198 + 0,020$, где $0 < \theta \le 1.$

Заметим, что точный корень уравнения (3.10) есть $\xi = 1,2.$

1.1.3 Метод касательной.

Его отличие от предыдущего метода состоит в том, что на *k*-й итерации вместо хорды проводится касательная к кривой *y = F(x)* при $x = c_{k-1}$ и ищется точка пересечения касательной с осью абсцисс. При этом не обязательно задавать отрезок [*a, b*], содержащий корень уравнения , а достаточно лишь найти некоторое начальное приближение корня $x = c_0$ (Рисунок 3.4).

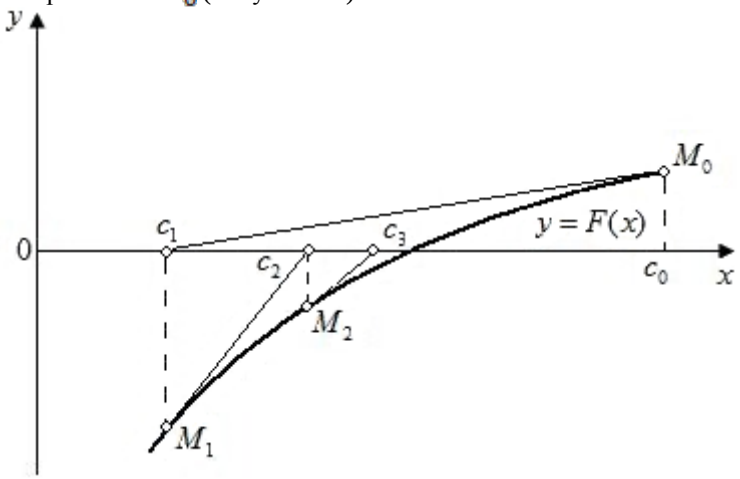

Рисунок 3.4

Уравнение касательной, проведенной к кривой *y = F(x)* в точке M_0 с координатами c_0 и $F(c_0)$, имеет вид

$$y - F(c_0) = F'(c_0)(x - c_0).$$

Отсюда найдем следующее приближение корня c_1 как абсциссу точки пересечения касательной с осью *x* (*y* = 0):

$$c_1 = c_0 - F(c_0)/F'(c_0).$$

Аналогично могут быть найдены и следующие приближения как точки пересечения с осью абсцисс касательных, проведенных в точках M_1, M_2 и т. д. Формула для *k*-го приближения имеет вид

$$c_k = c_{k-1} - F(c_{k-1})/F'(c_{k-1}), \qquad k = 1, 2, \ldots \tag{3.13}$$

При этом необходимо, чтобы $F'(c_{k-1})$ не равнялась нулю.

Из формулы следует, что на каждой итерации объем вычислений в методе Ньютона больший, чем в рассмотренных ранее методах, поскольку приходится находить значение не только функции *F(x),* но и ее производной. Однако скорость сходимости здесь значительно выше, чем в других методах.

Остановимся на некоторых вопросах, связанных со сходимостью метод Ньютона и его использованием. Имеет место следующая теорема**.**

Теорема. Пусть x=c - корень уравнения , т. е. F(c) =0, а F'(c)≠ 0 и F"(x) непрерывна. Тогда существует окрестность D корня c (c $\in$ D) такая, что если начальное приближение c_0 принадлежит этой окрестности, то для метода Ньютона последовательность значений $\{c_k\}$ сходится к c при $k \to \infty$. При этом для погрешности корня $\varepsilon_k = c - c_k$ имеет место соотношение

$$\lim_{k\to\infty} \left| \frac{\varepsilon_k}{\varepsilon_{k-1}^2} \right| = \left| \frac{F''(c)}{2F'(c)} \right|.$$

Фактически это означает, что на каждой итерации погрешность возводится в квадрат, т. е. число верных знаков корня удваивается. Если

$$\left| \frac{F''(c)}{2F'(c)} \right| \approx 1,$$

то легко показать, что при $|\varepsilon_0| \le 0.5$ пяти-шести итераций достаточно для получения минимально возможной погрешности при вычислениях с двойной точностью. Действительно, погрешность теоретически станет в этом случае величиной порядка 2^{-64}, что намного меньше, чем максимальная погрешность округления при вычислениях с двойной точностью, равная 2^{-53}. Заметим, что для получения столь малой погрешности в методе деления отрезка пополам потребовалось бы более 50 итераций.

ПРИМЕР. Для иллюстрации рассмотрим уравнение $x^2 - 0.25 = 0$ и найдем методом Ньютона один из его

корней, например $x = c = 0.5$. Для данного уравнения $F''(c)/F'(c) = 1$. Выберем $c_0 = 1$, тогда $\varepsilon_0 = -0.5$.
Проводя вычисления с двойной точностью, получим следующие значения погрешностей:

$\varepsilon_1 = -1.25 \cdot 10^{-1}$, $\varepsilon_3 = -1.52 \cdot 10^{-4}$, $\varepsilon_5 = -5.55 \cdot 10^{-16}$,

$\varepsilon_2 = -1.25 \cdot 10^{-2}$, $\varepsilon_4 = -2.32 \cdot 10^{-8}$, $\varepsilon_6 = 0$.

Таким образом, после шести итераций погрешность в рамках арифметики с двойной точностью исчезла.
Трудность в применении метода Ньютона состоит в выборе начального приближения, которое должно находиться в окрестности *D*. При неудачном выборе начального приближения итерации могут расходиться.
ПРИМЕР. Для уравнения arctg $x = 0$ (корень $x = c = 0$) при начальном приближении $c_0 = 1.5$ первые шесть итераций приводят к погрешностям

$\varepsilon_1 = 1.69$, $\varepsilon_3 = 5.11$, $\varepsilon_5 = 1.58 \cdot 10^3$,

$\varepsilon_2 = -2.32$, $\varepsilon_4 = -32.3$, $\varepsilon_6 = -3.89 \cdot 10^6$.

Очевидно, что итерации здесь расходятся.
Для предотвращения расходимости иногда целесообразно использовать смешанный алгоритм. Он состоит в том, что сначала применяется всегда сходящийся метод (например, метод деления отрезка пополам), а после некоторого числа итераций — быстро сходящийся метод Ньютона.

1.1.4 Метод простой итерации.

Для использования этого метода исходное нелинейное уравнение записывается в виде

$$x = f(x) \tag{3.14}$$

Пусть известно начальное приближение корня $x = c_0$. Подставляя это значение в правую часть уравнения (3.14), получаем новое приближение

$$c_1 = f(c_0).$$

Подставляя каждый раз новое значение корня в $x = f(x)$ (3.14), получаем последовательность значений

$$c_k = f(c_{k-1}), \qquad k = 1,2,\ldots$$

Итерационный процесс прекращается, если результаты двух последовательных итераций близки, т. е. если выполнено неравенство $|c_k - c_{k-1}| < \varepsilon$. Заметим, что в методе простой итерации для невязки, полученной на *k-й* итерации, выполнено соотношение

$$\gamma_k = c_k - f(c_k) = c_k - c_{k+1}.$$

Таким образом, условие малости невязки на *k-й* итерации оказывается эквивалентным условию близости k-го и $k+1$-го приближений.
Достаточное условие сходимости метода простой итерации дается следующей теоремой.
Теорема.
Пусть x=c - корень уравнения $x = f(x)$ (3.14), т. е. c= f(c), а $|f'| < 1$ и $f'(x)$ непрерывна. Тогда существует окрестность D корня c (c $\in$ D) такая, что если начальное приближение c_0 принадлежит этой окрестности, то для метода простой итерации последовательность значений $\{c_k\}$ сходится к c при к$\to \infty$.
Метод простой итерации рассмотрен нами для уравнения (3.14). К такому виду можно привести и более общее уравнение , аналогично тому, как это делалось при решении систем линейных уравнений:

$$\begin{aligned} &F(x) = 0 \\ &\tau F(x) = 0 \\ &x = x - \tau F(x) \end{aligned} \tag{3.15}$$

Здесь $\tau \neq 0$ - некоторое число. Уравнение (3.15 эквивалентно 3.14 функции $f(x) = x - \tau F(x)$. За счет выбора значения параметра τ можно добиваться сходимости метода простой итерации и повышения скорости сходимости. Например, если на некотором отрезке, содержащем корень уравнения, производная $F'(x)$ ограничена константами *m* и *M:*

$$0 < m < F'\{x) < M,$$

то для производной $f'(x)$ будет справедливо неравенство

$$1 - \tau M < f'(x) < 1 - \tau m.$$

Выбирая $\tau = 2/(M + m)$, получаем

$$-\frac{M-m}{M+m} < f'(x) < \frac{M-m}{M+m},$$

г. е. $|f'(x)| < 1$, что обеспечивает сходимость метода простой итерации.
Параметр τ в (3.15 можно выбирать и переменным, зависящим от

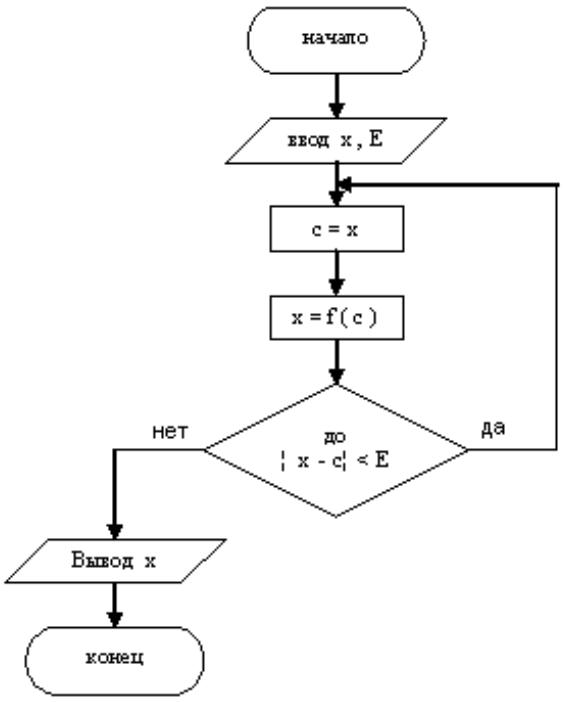

Рисунок 3.5

номера итерации. Так, если положить $\tau_k = 1/F'(c_{k-1})$, то метод простой итерации для уравнения (3.15 примет вид

$$c_k = c_{k-1} - F(c_{k-1})/F'(c_{k-1}).$$

Это соотношение совпадает с формулой метода Ньютона . Следовательно, метод Ньютона можно трактовать как частный случай метода простой итерации с переменным τ.

На рисунке 3.5 представлен алгоритм решения нелинейного уравнения 3.14 методом простой итерации. Здесь x - начальное приближение корня, а в дальнейшем — значение корня после каждой итерации, c - результат предыдущей итерации. В данном алгоритме предполагалось, что итерационный процесс сходится. Если такой уверенности нет, то необходимо ограничить число итераций и ввести для них счетчик.

§2. Действительные и комплексные корни алгебраических уравнений.

Действительные корни.

Рассмотренные выше методы решения нелинейных уравнений пригодны как для трансцендентных, так и для алгебраических уравнений. Вместе с тем при нахождении корней многочленов приходится сталкиваться с некоторыми особенностями. В частности, при рассмотрении точности вычислительного процесса отмечалась чувствительность к погрешностям значений корней многочлена. С другой стороны, по сравнению с трансцендентными функциями многочлены имеют то преимущество, что заранее известно число их корней.

Напомним некоторые известные из курса алгебры свойства алгебраических уравнений с действительными коэффициентами вида

$$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 = 0. \qquad (3.16)$$

1. Уравнение степени n имеет всего n корней с учетом кратности, среди вторых могут быть как действительные, так и комплексные.
2. Комплексные корни образуют комплексно - сопряженные пары, т. е. каждому корню $x = c + id$ соответствует корень $x = c - id$.

Одним из способов решения уравнения 3.16 является *метод понижения порядка.* Он состоит в том, что после нахождения какого - либо корня $x = c$ данное уравнение можно разделить на $x - c$, понизив его порядок до n — 1. Правда, при таком способе нужно помнить о точности, поскольку даже небольшая погрешность в значении первого корня может привести к накапливанию погрешности в дальнейших вычислениях.

Рассмотрим применение метода Ньютона к решению уравнения (3.16) в соответствии с формулой (3.13) итерационный процесс для нахождения корня нелинейного уравнения (3.14) имеет вид

$$x_k = x_{k-1} - \frac{F(x_{k-1})}{F'(x_{k-1})},$$

$$F(x) = a_0 + a_1 x + \ldots + a_n x^n, \qquad F'(x) = a_1 + 2a_2 x + \ldots + n a_n x^{n-1}.$$

Для вычисления значений многочленов *F(x)* и *F'(x)* в точке $x=x_{k-1}$ может быть использована схема Горнера.

Естественно, при использовании метода Ньютона должны выполняться условия сходимости. При их соблюдении в результате численного решения получается значение того корня, который находится вблизи заданного начального приближения x_0.

Заметим, что для уменьшения погрешностей лучше сначала находить меньшие по модулю корни многочлена и сразу удалять их из уравнения, приводя его к меньшей степени. Поэтому, если отсутствует информация о величинах корней, в качестве начальных приближений принимают числа 0, ±1 и т. д.

Комплексные корни.

При использовании компьютера имеется возможность работать с комплексными числами; поэтому изложенный метод Ньютона может быть использован (с необходимым обобщением) и для нахождения комплексных корней многочленов. При этом, если в качестве начального приближения x_0 взять комплексное число, то последующие приближения и окончательное значение корня могут оказаться комплексными. Ниже рассмотрим другой подход к отысканию комплексных корней.

Комплексные корни попарно сопряженные, и при их исключении порядок уравнения уменьшается на два, поскольку оно делится сразу на квадратный трехчлен, т. е.

$$F(x) = (x^2 + px + q)(b_n x^{n-2} + \cdots + b_2) + b_1 x + b_0 \qquad (3.17)$$

Линейный остаток $b_1 x + b_0$ равен нулю, если *p, q* выражаются с помощью найденных корней:

$$p = -2c, \quad q = c^2 + d^2, \quad x = c \pm id.$$

Представление (3.17) может быть также использовано для нахождения *p, q,* а значит, и для определения корней. Эта процедура лежит в основе *метода Лина.* Суть этого метода состоит в следующем. Предположим, что коэффициенты b_0, b_1 равны нулю. Тогда, сравнивая коэффициенты при одинаковых степенях *x* многочлена *F(x)* в выражениях (3.16) и (3.17), можно получить (для упрощения выкладок $b_n = a_n = 1$)

$$\begin{aligned} b_{n-1} &= a_{n-1} - p, \\ b_{n-2} &= a_{n-2} - p b_{n-1} - q \\ &\cdots\cdots\cdots \\ b_2 &= a_2 - p b_3 - p b_4 ; \end{aligned} \qquad (3.18)$$

$$p = (a_1 - \frac{q b_3)}{b_2}, \quad q = a_0 / b_2 \qquad (3.19)$$

В соотношения (3.19) входят коэффициенты b_2 и b_3 которые являются функциями *p* и *q.* Действительно, задав значения *p* и *q,* из соотношений (3.18) можно последовательно найти b_2 и b_3. Поэтому соотношения (3.19) представляют собой систему двух нелинейных уравнений относительно *p* и *q:*

$$\begin{aligned} p &= f_1(p, q), \\ q &= f_2(p, q). \end{aligned}$$

Такая система в методе Лина решается методом простой итерации: задаются начальные приближения для *p, q* которые используются для вычисления коэффициентов $b_{n-1}, b_{n-2}, \ldots, b_2,$ затем из уравнений (3.19) уточняются значения *p, q.* Итерационный процесс вычисления этих величин продолжается до тех пор, пока их изменения в двух последовательных итерациях не станут малыми.

Широко распространен также другой метод, основанный на выделении квадратичного множителя $x^2 + px + q$, — *метод Бэрстоу.* Он использует метод Ньютона для решения системы двух уравнений.

§3 Системы уравнений. Метод простой итерацию. Метод Ньютона.

3.1 Системы уравнений.

Вводные замечания.

Многие практические задачи сводятся к решению *системы нелинейных уравнений.*

Пусть для вычисления неизвестных $x_1, x_2, \ldots, x_n$ требуется решить систему *n* нелинейных уравнений

$$\begin{aligned} &F_1(x_1, x_2, \ldots, x_n) = 0, \\ &F_2(x_1, x_2, \ldots, x_n) = 0, \qquad (15) \\ &\cdots\cdots\cdots \\ &F_n(x_1, x_2, \ldots, x_n) = 0. \end{aligned} \qquad (3.20)$$

В векторной форме эту систему можно записать как

$$\mathbf{F(x) = O},$$

где

$$\mathbf{F}=\{F_1, F_2, \dots, F_n\}, \quad x=\{x_1, x_2, \dots, x_n\}.$$

В отличие от систем линейных уравнений не существует прямых методов решения нелинейных систем общего вида. Лишь в отдельных случаях систему (3.20) можно решить непосредственно. Например, для случая двух уравнений иногда удается выразить одно неизвестное через другое и таким образом свести задачу к решению одного нелинейного уравнения относительно одного неизвестного.

Для решения систем нелинейных уравнений обычно используются итерационные методы. Ниже будут рассмотрены некоторые из них: метод простой итерации, метод Зейделя и метод Ньютона.

3.2 Метод простой итерации.

Систему уравнений (3.20) представим в виде

$$\begin{aligned} x_1 &= f_1(x_1, x_2, \dots, x_n), \\ x_2 &= f_2(x_1, x_2, \dots, x_n), \\ &\dots\dots\dots\dots \\ x_n &= f_n(x_1, x_2, \dots, x_n) \end{aligned} \qquad (3.21)$$

Для решения этой системы можно использовать *метод простой итерации,* аналогичный соответствующему методу для одного уравнения. Значения неизвестных на *k-й* итерации будут найдены с использованием их значений на предыдущей итерации $x_1^{(k-1)}, x_2^{(k-1)}, \dots, x_n^{(k-1)}$ как

$$x_i^{(k)} = f_i(x_1^{(k-1)}, x_2^{(k-1)}, \dots, x_n^{(k-1)}), i = 1,2,\dots,n \qquad (3.22)$$

Систему (3.21) можно решать и *методом Зейденя,* напоминающим метод Гаусса-Зейделя решения систем линейных уравнений. Значение $x_i^{(k)}$ находится из *i*-го уравнения системы (3.21) с использованием уже вычисленных на текущей итерации значений неизвестных. Таким образом, значения неизвестных на *k-й.* итерации будут находиться не с помощью (3.22), а с помощью соотношения

$$x_i^{(k)} = f_i(x_1^{(k)}, \dots, x_{i-1}^{(k)}, x_i^{(k-1)}, \dots, x_n^{(k-1)}), i = 1,2,\dots,n$$

Итерационный процесс в обоих методах продолжается до тех пор, пока изменения всех неизвестных в двух последовательных итерациях не станут малыми.

При использовании метода простой итерации и метода Зейделя успех во многом определяется удачным выбором начальных приближений неизвестных: они должны быть достаточно близкими к истинному решению. В Противном случае итерационный процесс может не сойтись.

3.3 Метод Ньютона.

Этот метод обладает гораздо более быстрой сходимостью, чем метод простой итерации и метод Зейделя. В случае одного уравнения *F(x)*= 0 алгоритм метода Ньютона был легко получен путем записи уравнения касательной к кривой *y =F(x).* По сути для нахождения нового приближения функция *F(x)* заменялась линейной функцией, т. е. раскладывалась в ряд Тейлора, при этом член, содержащий вторую производную, отбрасывался (как и все последующие члены). Та же идея лежит в основе метода Ньютона для системы уравнений: функции $F_i(x_1, x_2, \dots, x_n)$ раскладываются в ряд Тейлора, причем в разложении отбрасываются члены, содержащие вторые (и более высоких порядков) производные.

Пусть приближенные значения неизвестных системы (3.20), полученные на предыдущей итерации, равны соответственно $x_1^{(k-1)}, x_2^{(k-1)}, \dots, x_n^{(k-1)}$. Задача состоит в нахождении приращений (поправок) к этим значениям $\Delta x_1, \Delta x_2, \dots, \Delta x_n$, благодаря которым следующее приближение к решению системы (3.20) запишется в виде

$$\begin{aligned} x_1^{(k)} &= x_1^{(k-1)} + \Delta x_1, \\ x_2^{(k)} &= x_2^{(k-1)} + \Delta x_2, \\ &\dots \\ x_n^{(k)} &= x_n^{(k-1)} + \Delta x_n, \end{aligned} \qquad (3.23)$$

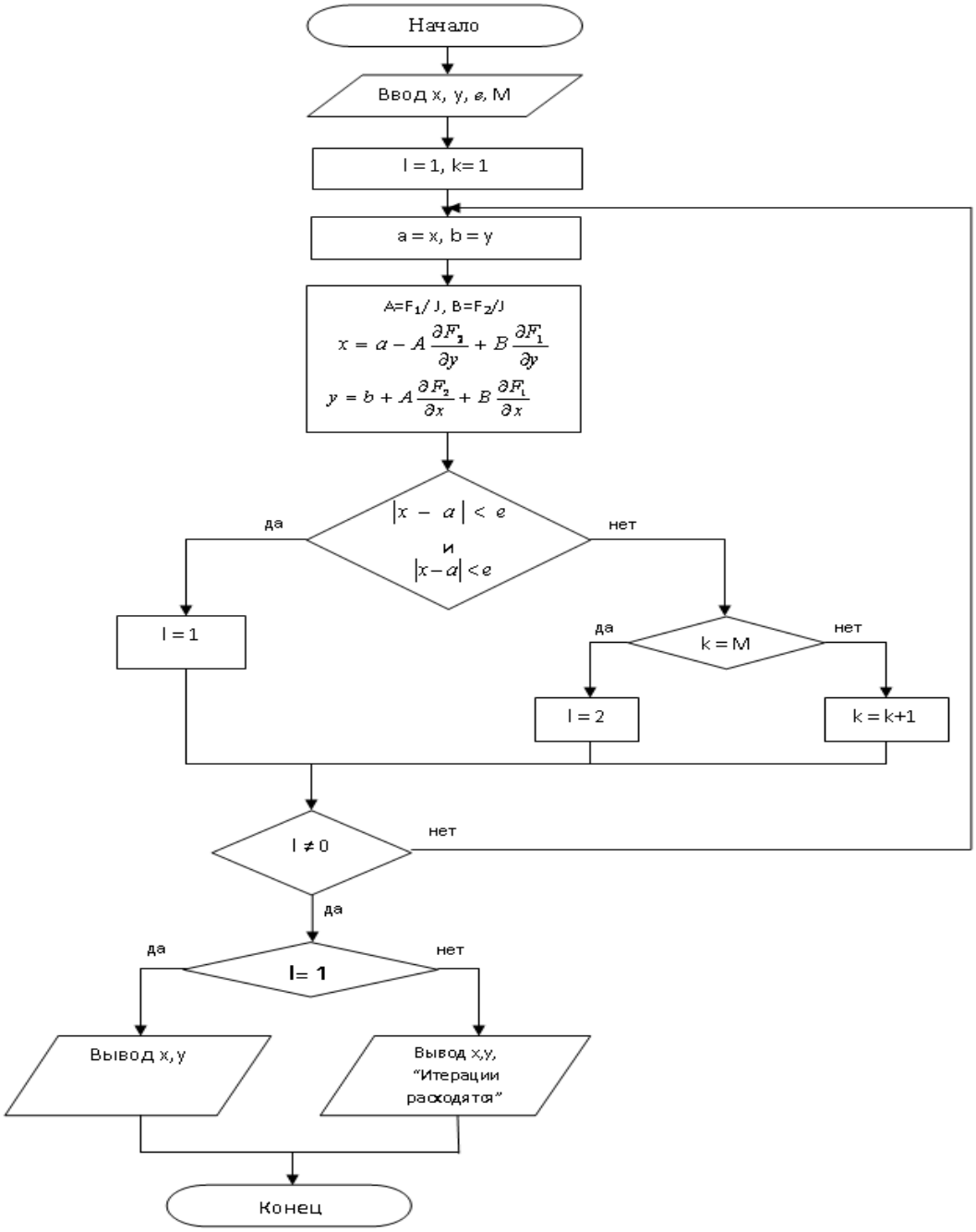

Рисунок 3.6

Проведем разложение левых частей уравнений (3.20) в ряд Тейлора, ограничиваясь лишь линейными членами относительно приращений:

$$
\begin{aligned}
F_1(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}) &\approx F_1 + \frac{\partial F_1}{\partial x_1}\Delta x_1 + \ldots + \frac{\partial F_1}{\partial x_n}\Delta x_n, \\
F_2(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}) &\approx F_2 + \frac{\partial F_2}{\partial x_1}\Delta x_1 + \ldots + \frac{\partial F_2}{\partial x_n}\Delta x_n, \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
F_n(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}) &\approx F_n + \frac{\partial F_n}{\partial x_1}\Delta x_1 + \ldots + \frac{\partial F_n}{\partial x_n}\Delta x_n.
\end{aligned}
\qquad (3.24)
$$

В правых частях этих соотношений значения $F_1, F_2, \ldots, F_n$ и их производных вычисляются в точке $x^{(k-1)} = (x_1^{(k-1)}, x_2^{(k-1)}, \ldots, x_n^{(k-1)})$

Поскольку в соответствии с (3.20) левые части (3.24) должны обращаться в нуль, приравняем нулю и правые части, т. е. найдем новое приближение из условия равенства нулю разложений функций F_i. Получим следующую систему линейных алгебраических уравнений относительно приращений:

$$\frac{\partial F_1}{\partial x_1}\Delta x_1 + \frac{\partial F_1}{\partial x_2}\Delta x_2 + \ldots + \frac{\partial F_1}{\partial x_n}\Delta x_n = -F_1,$$
$$\frac{\partial F_2}{\partial x_1}\Delta x_1 + \frac{\partial F_2}{\partial x_2}\Delta x_2 + \ldots + \frac{\partial F_2}{\partial x_n}\Delta x_n = -F_2,$$
$$. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad .$$
$$\frac{\partial F_n}{\partial x_1}\Delta x_1 + \frac{\partial F_n}{\partial x_2}\Delta x_2 + \ldots + \frac{\partial F_n}{\partial x_n}\Delta x_n = -F_n, \qquad (3.25)$$

Определителем системы (3.25) является *якобиан.*
Для существования единственного решения системы (3.25) он должен быть отличным от нуля на каждой итерации.
Таким образом, итерационный процесс решения системы уравнений (3.20) методом Ньютона состоит в определении приращений $\Delta x_1, \Delta x_2, \ldots , \Delta x_n$ к значениям неизвестных на каждой итерации посредством решения системы (3.25). В методе Ньютона также важен удачный выбор начального приближения для обеспечения хорошей сходимости. Сходимость ухудшается с увеличением числа уравнений системы. В качестве примера рассмотрим использование метода Ньютона для решения системы двух уравнений

$$F_1(x, y) = 0, F_1(x, y) = 0 \qquad (3.26)$$

Пусть приближенные значения неизвестных равны *a, b*. Предположим, что якобиан системы (3.26) при $x=a$, $y=b$ отличен от нуля, т. е.

$$J = \begin{vmatrix} \frac{\partial F_1}{\partial x} & \frac{\partial F_1}{\partial y} \\ \frac{\partial F_2}{\partial x} & \frac{\partial F_2}{\partial y} \end{vmatrix} \neq 0$$

Тогда следующие приближения неизвестных можно записать в виде

$$x = a - \frac{1}{J}\left(F_1 \frac{\partial F_2}{\partial y} - F_2 \frac{\partial F_1}{\partial y}\right),$$
$$y = b + \frac{1}{J}\left(F_1 \frac{\partial F_2}{\partial x} - F_2 \frac{\partial F_1}{\partial x}\right).$$

Величины, стоящие в правых частях, вычисляются при $x= a$, $y= b$.
Алгоритм метода Ньютона для решения системы двух уравнений изображен на рис. 3.6. В качестве исходных данных задаются начальные приближения неизвестных *x, y,* погрешность ε и допустимое число итераций *M.* В условной конструкции, проверяющей выполнение критерия завершения итерации, используется логическая операция «и», имеющаяся в современных языках программирования. Если итерации сойдутся, то выводятся значения *x, y;* в противном случае — текущие значения *x, y* и соответствующее сообщение.

Глава 4
РЕШЕНИЕ СИСТЕМ ЛИНЕЙНЫХ УРАВНЕНИЙ

Линейные системы. К решению систем линейных уравнений сводятся многочисленные практические задачи. Можно с полным основанием утверждать, что решение линейных систем является одной из самых распространенных и важных задач вычислительной математики.

Запишем *систему n линейных алгебраических уравнений с n неизвестными:*

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

Совокупность коэффициентов этой системы запишем в виде таблицы:

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix}.$$

Данная таблица n^2 элементов, состоящая из n строк и n столбцов, называется *квадратной матрицей* порядка . Если подобная таблица содержит mn элементов, расположенных в m строках и n столбцах, то она называется *прямоугольной матрицей.*

Используя понятие матрицы A, систему уравнений (5.1) можно записать в матричном виде:

$$Ax = B \qquad (4.2)$$

где x *и* b — вектор-столбец неизвестных и вектор-столбец правых частей соответственно:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix}, \qquad b = \begin{pmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ b_n \end{pmatrix},$$

или, в более компактной записи,

$$x = \{x_1, x_2, \ldots x_n\}., \qquad b = \{b_1, b_2, \ldots b_n\}.$$

В ряде случаев получаются системы уравнений с некоторыми специальными видами матриц. Вот некоторые примеры таких матриц:

$$A=\begin{pmatrix} 2 & 1 & -1 \\ 1 & 3 & 2 \\ -1 & 2 & 4 \end{pmatrix}, \qquad B=\begin{pmatrix} 1 & 2 & 3 \\ 0 & -1 & 1 \\ 0 & 0 & 2 \end{pmatrix},$$

$$C=\begin{pmatrix} 1 & 2 & 1 & 0 & 0 & 0 \\ 2 & -1 & 2 & 0 & 0 & 0 \\ 3 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & -1 & 1 \\ 0 & 0 & 0 & -1 & 4 & -1 \\ 0 & 0 & 0 & 2 & 1 & 1 \end{pmatrix}, \qquad F=\begin{pmatrix} 3 & 2 & 0 & 0 & 0 & 0 \\ 1 & 2 & -1 & 0 & 0 & 0 \\ 0 & 3 & -2 & 2 & 0 & 0 \\ 0 & 0 & 1 & 2 & -1 & 0 \\ 0 & 0 & 0 & 1 & 3 & 1 \\ 0 & 0 & 0 & 0 & -1 & 3 \end{pmatrix},$$

$$E=\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad O=\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Здесь:

A — симметрическая матрица (ее элементы расположены симметрично относительно главной диагонали ($a_{ij}=a_{ji}$));

B — верхняя треугольная матрица с равными нулю элементами, расположенными ниже диагонали;

C — клеточная матрица (ее ненулевые элементы составляют отдельные группы (клетки));

F — ленточная матрица (ее ненулевые элементы составляют «ленту», параллельную диагонали (в данном случае ленточная матрица *D* одновременно является тфакже трехдиагональной));

E — единичная матрица (частный случай диагональной);

O — нулевая матрица.

Определителем (детерминантом) матрицы *A* n-го порядка называется число *D*, равное

$$D=\det A=\begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix}=\sum(-1)^k a_{1\alpha}a_{2\beta}\dots a_{n\omega}. \qquad (4.3)$$

Здесь индексы $\alpha,\beta,\dots,\omega$ пробегают все возможные n! перестановок номеров 1, 2, ..., n; k — число инверсий в данной перестановке.

Необходимым и достаточным условием существования единственного решения системы линейных уравнений является условие $D\neq 0$. В случае равенства нулю определителя системы матрица называется *вырожденной*, при этом система линейных уравнений (5.1) либо не имеет решения, либо имеет их бесчисленное множество.

Все эти случаи легко проиллюстрировать геометрически для системы

$$\begin{aligned} a_1x+b_1y&=c_1, \\ a_2x+b_2y&=c_2. \end{aligned} \qquad (4.4)$$

Каждое уравнение описывает прямую на плоскости; координаты точки пересечения указанных прямых являются решением системы (5.4).

Рассмотрим три возможных случая взаимного расположения двух прямых на плоскости:

1) прямые пересекаются — коэффициенты системы (5.4) не пропорциональны:

$$\frac{a_1}{a_2}\neq\frac{b_1}{b_2}; \qquad (4.5)$$

2) прямые параллельны — коэффициенты системы (5.4) подчиняются условиям

$$\frac{a_1}{a_2}=\frac{b_1}{b_2}\neq\frac{c_1}{c_2}; \qquad (4.6)$$

3) прямые совпадают — все коэффициенты (5.4) пропорциональны:

$$\frac{a_1}{a_2} = \frac{b_1}{b_2} = \frac{c_1}{c_2}. \qquad (4.7)$$

Запишем определитель D системы (5.4) в виде

$$D = \begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}.$$

Отметим, что при выполнении условия (5.5) $D \neq 0$, и система (5.4) имеет единственное решение. В случаях отсутствия решения или при бесчисленном множестве решений имеют место соответственно соотношения (5.6) или (5.7), из которых получаем $D=0$.

На практике, особенно при вычислениях на ЭВМ, когда происходят округление или отбрасывание младших разрядов чисел, далеко не всегда удается получить точное равенство определителя нулю. При $D \approx 0$ прямые могут оказаться почти параллельными (в случае системы двух уравнений); координаты точки пересечения этих прямых весьма чувствительны к изменению коэффициентов системы.

Таким образом, малые погрешности вычислений пли исходных данных могут привести к существенным погрешностям в решении. Такие системы уравнений называются *плохо обусловленными.*

Заметим, что условие $D \approx 0$ является необходимым для плохой обусловленности системы линейных уравнений, но не достаточным. Например, система уравнений n-го порядка с диагольной матрицей с элементами $a_{ii}=0.1$ не является плохо обусловленной, хотя ее определитель мал ($D = 10^{-n}$).

Геометрическая иллюстрация системы двух уравнений; при малом изменении параметров одной из прямых координаты точки пересечения мало изменяются в случае *a* и заметно изменяются в случае *b*

Приведенные соображения справедливы и для любого числа уравнений системы (5.1) хотя в случае $n>3$ нельзя привести простые геометрические иллюстрации. При $n = 3$ каждое уравнение описывает плоскость в пространстве, и в случае почти параллельных плоскостей или линий их опарного пересечения получаем плохо обусловленную систему трех уравнений.

О методах решения линейных систем. Методы решения систем линейных уравнений делятся на две группы — прямые и итерационные. *Прямые методы* используют конечные соотношения (формулы) для вычисления неизвестных. Они дают решение после выполнения заранее известного числа операций. Эти методы сравнительно просты и наиболее универсальны, т. е. пригодны для решения широкого класса линейных систем.

Вместе с тем прямые методы имеют и ряд недостатков. Как правило, они требуют хранения в оперативной памяти компьютера сразу всей матрицы, и при больших значениях *n* расходуется много места в памяти. Далее, прямые методы обычно не учитывают структуру матрицы — при большом числе нулевых элементов в разреженных матрицах (например, клеточных или ленточных) эти элементы занимают место в памяти машины, и над ними проводятся арифметические действия. Существенным недостатком прямых методов является также накапливание погрешностей в процессе решения, поскольку вычисления на любом этапе используют результаты предыдущих операций. Это особенно опасно для больших систем, когда резко возрастает общее число операций, а также для плохо обусловленных систем, весьма чувствительных к погрешностям. В связи с этим прямые методы используются обычно для сравнительно небольших ($n \leq 200$) систем с плотно заполненной матрицей и не близким к нулю определителем.

Отметим еще, что прямые методы решения линейных систем иногда называют *точными,* поскольку решение выражается в виде точных формул через коэффициенты системы. Однако точное решение может быть получено лишь при выполнении вычислений с бесконечным числом разрядов (разумеется, при точных значениях коэффициентов системы). На практике при использовании ЭВМ вычисления проводятся с ограниченным числом знаков, определяемым разрядностью машины. По этому неизбежны погрешности в окончательных результатах.

Итерационные методы — это методы последовательных приближений. В них необходимо задать некоторое приближенное решение — *начальное приближение.* После этого с помощью некоторого алгоритма проводится один цикл вычислений, называемый *итерацией.* В результате итерации находят новое приближение. Итерации проводятся до получения решения с требуемой точностью. Алгоритмы решения линейных систем с использованием итерационных методов обычно более сложные по сравнению с прямыми методами. Объем вычислений заранее определить трудно.

Тем не менее итерационные методы в ряде случаев предпочтительнее. Они требуют хранения в памяти машины не всей матрицы системы, а лишь нескольких векторов с *n* компонентами. Иногда элементы матрицы можно совсем не хранить, а вычислять их по мере необходимости. Погрешности окончательных результатов при использовании итерационных методов не накапливаются, поскольку точность вычислений в каждой итерации определяется лишь результатами предыдущей итерации и практически не зависит от ранее выполненных вычислений. Эти достоинства итерационных методов делают их особенно полезными в случае большого числа уравнений, а также плохо обусловленных систем. Следует отметить, что при этом сходимость итераций может быть очень медленной; поэтому ищутся эффективные пути ее ускорения.

Итерационные методы могут использоваться для уточнения решений, полученных с помощью прямых методов. Такие смешанные алгоритмы обычно довольно эффективны, особенно для плохо обусловленных систем. В последнем случае могут также применяться методы регуляризации.

Другие задачи линейной алгебры. Кроме решения систем линейных уравнений существуют другие задачи линейной алгебры — вычисление определителя, обратной матрицы, собственных значений матрицы и др.
Легко вычисляются лишь определители невысоких порядков и некоторые специальные типы определителей. В частности, для определителей второго и третьего порядков соответственно имеем

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

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{21}a_{32}a_{13} - a_{31}a_{22}a_{13} - a_{21}a_{12}a_{33} - a_{32}a_{23}a_{11}.$$

Определитель треугольной матрицы равен произведению ее элементов, расположенных на главной диагонали: $D=a_{11}a_{22}\ldots a_{nn}$. Отсюда также следует, что определитель единичной матрицы равен единице, а нулевой — нулю: detE=1, det O=0.
В общем случае вычисление определителя оказывается значительно более трудоемким. Определитель *D* порядка n имеет вид (5.3)

$$D = \sum (-1)^k a_{1\alpha}a_{2\beta}\ldots a_{n\omega}.$$

Из этого выражения следует, что определитель равен сумме *n*\ слагаемых, каждое из которых является произведением *n* элементов. Поэтому для вычисления определителя порядка *n* (без использования специальных приемов) требуется *(n*— 1)*n*! умножений и *n!* — 1 сложений, т. е. общее число арифметических операций равно

$$N = n \cdot n! - 1 \approx n \cdot n! \qquad (4.8)$$

Оценим значения *N* в зависимости от порядка *n* определителя:

n	3	10	20
N	17	$3.6 \cdot 10^7$	$5 \cdot 10^{19}$

Можно подсчитать время вычисления таких определителей на компьютере с заданным быстродействием. Примем для определенности среднее быстродействие равным 10 млн. операций в секунду. Тогда для вычисления определителя 10-го порядка потребуется около 3.6 сек, а при *n* = 20—свыше 150 тыс. лет.
Приведенные оценки указывают на необходимость разработки и использования экономичных численных методов, позволяющих эффективно проводить вычисления определителей.
Матрица $\boldsymbol{A}^{-1}$ называется, *обратной* по отношению к квадратной матрице *A,* если их произведение равно единичной матрице: $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = E$. В линейной алгебре доказывается, что всякая невырожденная матрица *A* (т.е. с отличным от нуля определителем *D)* имеет обратную. При этом

$$\det A^{-1} = 1/D.$$

Запишем исходную матрицу в виде

$$A = \begin{pmatrix} a_{11} & \ldots & a_{1j} & \ldots & a_{1n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{i1} & \ldots & a_{ij} & \ldots & a_{in} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n1} & \ldots & a_{nj} & \ldots & a_{nn} \end{pmatrix}.$$

Минором элемента a_{ij} называется определитель (n-1)-го порядка, образованный из определителя матрицы *A* зачеркиванием i-й строки и j-го столбца.
Алгебраическим дополнением A_{ij} элемента a_{ij} называется его минор, взятый со знаком плюс, если сумма *i+j* номеров строки *i* и столбца *j* четная, и со знаком минус, если эта сумма нечетная, т. е.

$$A_{ij} = (-1)^{i+j} \begin{vmatrix} a_{11} & \dots & a_{1,j-1} & a_{1,j+1} & \dots & a_{1n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{i-1,1} & \dots & a_{i-1,j-1} & a_{i-1,j+1} & \dots & a_{i-1,n} \\ a_{i+1,1} & \dots & a_{i+1,j-1} & a_{i+1,j+1} & \dots & a_{i+1,n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & \dots & a_{n,j-1} & a_{n,j+1} & \dots & a_{nn} \end{vmatrix}.$$

Каждый элемент z_{ij} (i,j = l, ..., *n)* обратной матрицы Z=A^{-1} равен отношению алгебраического дополнения A_{ij} элемента a_{ii} (не a_{ii}) исходной матрицы *A* к значению ее определителя *D:*

$$Z = A^{-1} = \begin{pmatrix} \frac{A_{11}}{D} & \frac{A_{21}}{D} & \dots & \frac{A_{n1}}{D} \\ \frac{A_{12}}{D} & \frac{A_{22}}{D} & \dots & \frac{A_{n2}}{D} \\ \dots & \dots & \dots & \dots \\ \frac{A_{1n}}{D} & \frac{A_{2n}}{D} & \dots & \frac{A_{nn}}{D} \end{pmatrix}$$

Здесь, как и выше, можно также подсчитать число операций, необходимое для вычисления обратной матрицы без использования специальных методов. Это число равно сумме числа операций, с помощью которых вычисляются n^2 алгебраических дополнений, каждое из которых является определителем (*n*-1)-го порядка, и n^2делений алгебраических до получений на определитель *D.* Таким образом, общее число операций для вычислений обратной матрицы равно

$$N = \left[(n-1)\cdot(n-1)!-1\right]n^2 + n^2 + n\cdot n!-1 = n^2\cdot n!-1.$$

Важной задачей линейной алгебры является также вычисление собственных значений матрицы.

§1. Прямые методы. Метод Гаусса. Метод главных диагоналей. Определитель и обратная матрица. Метод прогонки.

1.1 Прямые методы

Вводные замечания. Одним из способов решения системы линейных уравнений является *правило Крамера,* согласно которому каждое неизвестное представляется в виде отношения определителей. Запишем его для системы

$$a_1x + b_1y = c_1$$
$$a_2x + b_2y = c_2$$

Тогда

$$x = D_1 / D,\ \ y = D_1 / D,$$

$$D = \begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}, \quad D_1 = \begin{vmatrix} c_1 & b_1 \\ c_2 & b_2 \end{vmatrix}, \quad D_2 = \begin{vmatrix} a_1 & c_1 \\ a_2 & c_2 \end{vmatrix}.$$

Можно попытаться использовать это правило для решения систем уравнений произвольного порядка. Однако при большом числе уравнений потребуется выполнить огромное число арифметических операций, поскольку для вычислений *n* неизвестных необходимо найти значения определителей, число которых n+1. Количество арифметических операций можно оценить с учетом формулы (8). При этом предполагаем, что определители вычисляются непосредственно — без использования экономичных методов. Тогда получим

$$N = (n+1)(n\cdot n!-1) + n.$$

Поэтому правило Крамера молено использовать лишь для решения систем, состоящих из нескольких уравнений.

Известен также метод решения линейной системы с использованием обратной матрицы. Система записывается в виде $Ax = b$. Тогда, умножая обе части этого векторного уравнения слева на обратную матрицу A^{-1}, получаем x= A^{-1}b. Однако если не использовать экономичных схем для вычисления обратной матрицы, этот способ также непригоден для практического решения линейных систем при больших значениях *n* из-за большого объема вычислений.

Наиболее распространенными среди прямых методов являются метод исключения Гаусса и его модификации.

Ниже рассматривается применение метода исключения для решения систем линейных уравнений, а также для вычисления определителя и нахождения обратной матрицы.

1.2 Метод Гаусса.

Он основан на приведении матрицы системы к треугольному виду. Это достигается последовательным исключением неизвестных из уравнений системы. Сначала с помощью первого уравнения исключается x_i из всех последующих уравнений системы. Затем с помощью второго уравнения исключается x_2 из третьего и всех последующих уравнений. Этот процесс, называемый *прямым ходом метода Гаусса,* продолжается до тех пор, пока в левой части последнего (n-го) уравнения не останется лишь один член с неизвестным x_n, т. е. матрица системы будет приведена к треугольному виду. (Заметим, что к такому виду приводится лишь невырожденная матрица, в противном случае метод Гаусса неприменим).

Обратный ход метода Гаусса состоит в последовательном вычислении искомых неизвестных: решая последнее сравнение, находим единственное неизвестное x_n. Далее, используя это значение, из предыдущего уравнения вычисляем x_{n-1} и т. д. Последним найдем x_1 из первого уравнения.

Рассмотрим применение метода Гаусса для системы

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1, \\ a_{21}x_1 + a_{22}x_2 + a_{33}x_3 &= b_2, \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3. \end{aligned} \tag{4.9}$$

Для исключения x_1 из второго уравнения прибавим к нему первое, умноженное на $-a_{21}/\ a_{11}$. Затем, умножив первое уравнение на $-a_{31}/\ a_{11}$ и прибавив результат к третьему уравнению, также исключаем из него x_1. Получив равносильную систему уравнений вида

$$\begin{aligned} &a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1, \\ &a'_{22}x_2 + a'_{23}x_2 = b'_2, \\ &a'_{32}x_2 + a'_{33}x_2 = b'_3; \end{aligned}$$

$$a'_{ij} = a_{ij} - \frac{a_{i1}}{a_{11}}a_{1j}, \quad i, j = 2,3, \tag{4.10}$$

$$b'_i = b_i - \frac{a_{i1}}{a_{11}}b_1, \qquad i = 2,3.$$

Теперь из третьего уравнения системы (5.10) нужно исключить x_2. Для этого умножим второе уравнение на $-a'_{32}/\ a'_{22}$ и прибавим результат к третьему. Получим

$$\begin{aligned} &a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1, \\ &a'_{22}x_2 + a'_{23}x_3 = b'_2, \\ &a''_{33}x_3 = b''_3; \\ &a''_{33} = a'_{33} - \frac{a'_{32}}{a'_{22}}a'_{23},\ b''_3 = b'_3 - \frac{a'_{32}}{a'_{22}}b'_2. \end{aligned} \tag{4.11}$$

Матрица системы (5.11) имеет треугольный вид. На этом заканчивается прямой ход метода Гаусса.

Заметим, что в процессе исключения неизвестных приходится выполнять операции деления на коэффициенты a_{11}, a'_{22} и т. д. Поэтому они должны быть отличными от нуля; в противном случае необходимо соответственным образом переставить уравнения системы. Перестановка уравнений должна быть предусмотрена в вычислительном алгоритме при его реализации на компьютере.

Обратный ход начинается с решения третьего уравнения системы (5.11)

$$x_3 = b''_3 / a''_3.$$

Используя это значение, можно найти x_2 из второго уравнения, а затем x_1 из первого:

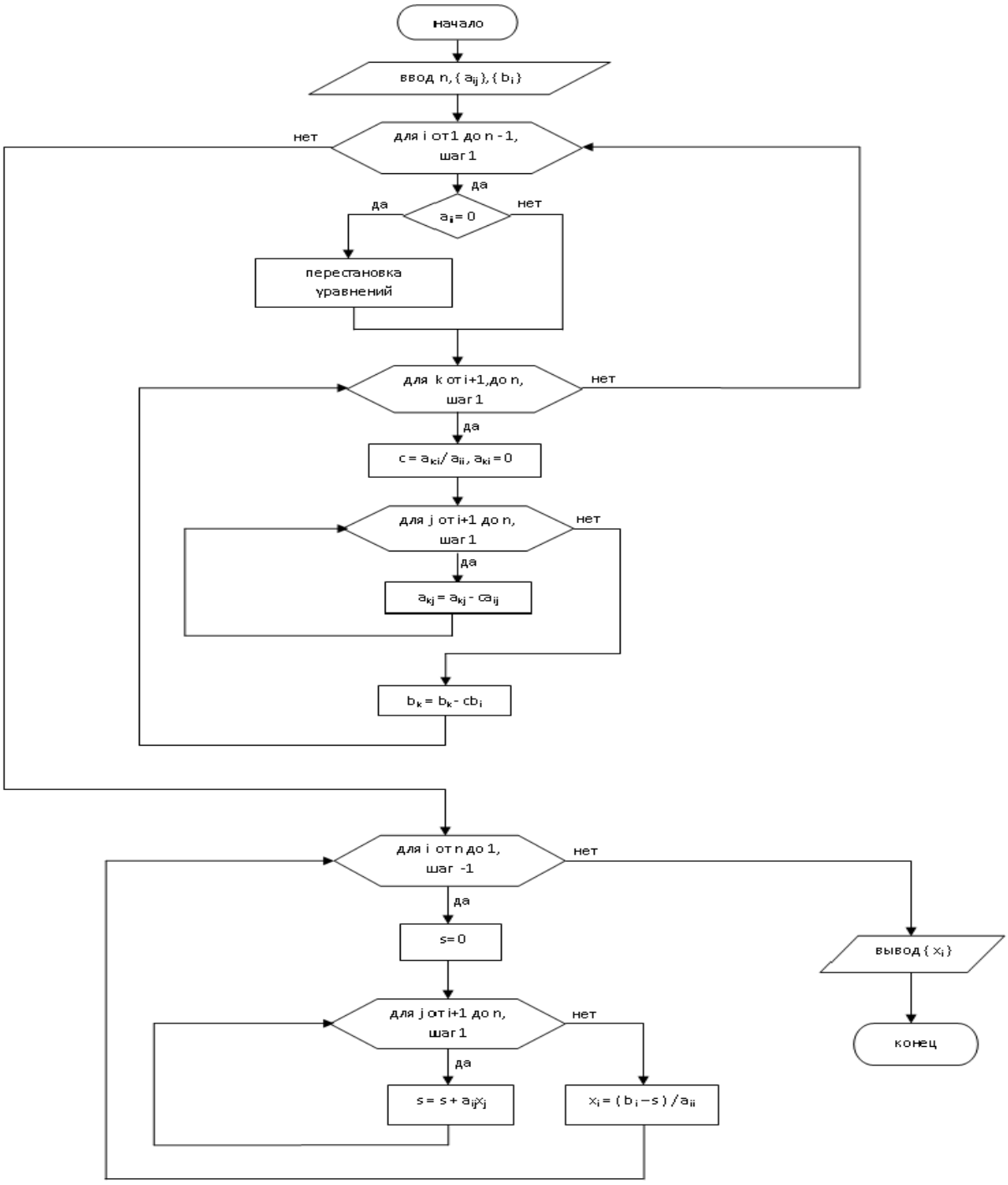

Рисунок 4.1 !!!!!

$$x_2 = \frac{1}{a_{22}^{'}}(b_2^{'} - a_{23}^{'}x_3),$$

$$x_1 = \frac{1}{a_{11}}(b_1 - a_{12}x_2 - a_{13}x_3).$$

Аналогично строится вычислительный алгоритм для линейной системы с произвольным числом уравнений.

Левая часть блок-схемы соответствует прямому ходу. Поясним смысл индексов: i — номер уравнения, из которого исключается неизвестное x_k; j-номер столбца; k — номер неизвестного, которое исключается из оставшихся n-k уравнений (а также номер того уравнения, с помощью которого исключается из оставшихся n-k уравнений). Операция перестановки уравнений (т. е. перестановки соответствующих коэффициентов) служит для предотвращения деления на нулевой элемент. Правая часть блок-схемы описывает процесс обратного хода. Здесь i - номер неизвестного, которое определяется из i-го уравнения; j = i + 1, i + 2, ...-

номера уже найденных неизвестных.
Одной из модификаций метода Гаусса является *схема с выбором главного элемента*. Она состоит в том, что требование неравенства нулю диагональных элементов a_{kk}, на которые происходит деление в процессе исключения, заменяется более жестким: из всех оставшихся в k-м столбце элементов нужно выбрать наибольший по модулю и переставить уравнения так, чтобы этот элемент оказался на месте элемента a_{kk}.

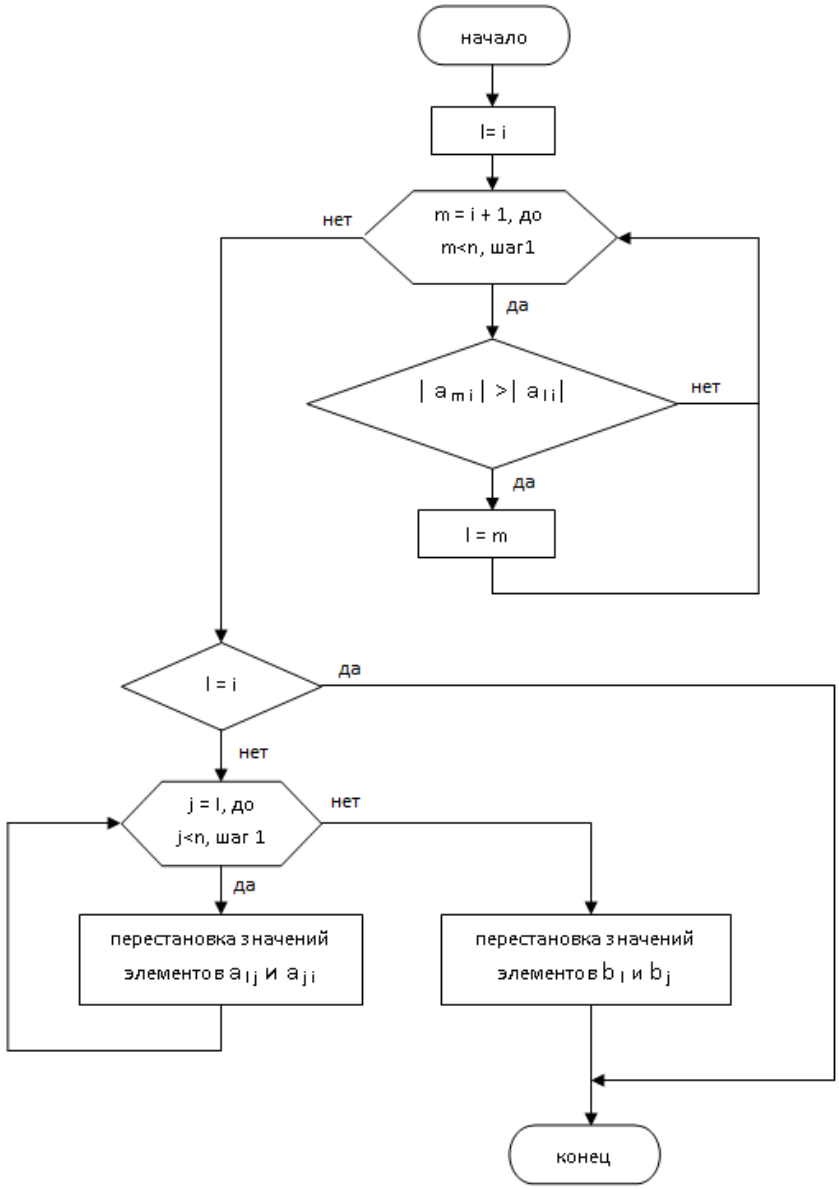

Рисунок 4.2 !!!!!

Блок-схема алгоритма выбора главного элемента приведена на Рисунок 4.1 . Она дополняет блок-схему метода Гаусса.
Здесь введены новые индексы: *l*— помер наибольшего по абсолютной величине элемента матрицы в столбце с номером *k* (т. е. среди элементов a_{kk}, …, a_{km},…, a_{kn}); m — текущий номер элемента, с которым происходит сравнение. Заметим, что диагональные элементы матрицы называются *ведущими* элементами; ведущий элемент a_{kk} — это коэффициент при *k-м* неизвестном в k-м уравнении на *k*-м шаге исключения.
Благодаря выбору наибольшего по модулю ведущего элемента уменьшаются множители, используемые для преобразовании уравнений, что способствует снижению погрешностей вычислений. Поэтому метод Гаусса с выбором главного элемента обеспечивает приемлемую точность решения для сравнительно небольшого числа (n< 100) уравнений. И только для плохо обусловленных систем решения, полученные по этому методу, ненадежны.
Метод Гаусса целесообразно использовать для решения систем с плотно заполненной матрицей. Все элементы матрицы и правые части системы уравнений находятся в оперативной памяти машины. Объем вычислений определяется порядком системы n: число арифметических операций примерно равно (2/3) n^3.

Пример. Рассмотрим алгоритм решения линейной системы методом Гаусса и некоторые особенности этого метода для случая трех уравнений:

$$\begin{aligned} 10x_1 - 7x_2 &= 7, \\ -3x_1 + 3x_2 + 6x_3 &= 4, \\ 5x_1 - x_2 + 5x_3 &= 6. \end{aligned}$$

Исключим x_1 из второго и третьего уравнений. Для этого сначала умножим первое уравнение на 0.3 и результат прибавим ко второму, а затем умножим первое же уравнение на -0.5 и результат прибавим к третьему. Получим

$$\begin{aligned} 10x_1 - 7x_2 &= 7, \\ -0.1x_2 + 6x_3 &= 6.1, \\ 2.5x_2 + 5x_3 &= 2.5. \end{aligned}$$

Прежде чем исключать x_2 из третьего уравнения, заметим, что коэффициент при x_2 во втором уравнении (ведущий элемент) мал; поэтому было бы лучше переставить второе и третье уравнения. Однако мы проводим сейчас вычисления в рамках точной арифметики и погрешности округлении не опасны, поэтому продолжим исключение. Умножим второе уравнение на 25 и результат сложим с третьим уравнением. Получим систему в треугольном виде:

$$\begin{aligned} 10x_1 - 7x_2 &= 7, \\ -0.1x_2 + 6x_3 &= 6.1, \\ 155x_3 &= 155. \end{aligned}$$

На этом заканчивается прямой ход метода Гаусса.

Обратный ход состоит в последовательном вычислении x_3, x_2, x_1 соответственно из третьего, второго, первого уравнений. Проведем эти вычисления:

$$x_3 = \frac{155}{155} = 1, \quad x_2 = \frac{6x_3 - 6.1}{0.1} = -1, \quad x_1 = \frac{7x_2 + 7}{10} = 0.$$

Подстановкой в исходную систему легко убедиться, что (0, - 1, 1) и есть ее решение.

Изменим теперь слегка коэффициенты системы таким образом, чтобы сохранить прежним решение и вместе с тем при вычислениях использовать округления. Таким условиям, в частности, соответствует система

$$\begin{aligned} 10x_1 - 7x_2 &= 7, \\ -3x_1 + 2.099x_2 + 6x_3 &= 3.901, \\ 5x_1 - x_2 + 5x_3 &= 6. \end{aligned}$$

Здесь изменены коэффициент при x_2 и правая часть второго уравнения. Будем снова вести процесс исключения, причем вычисления проведем в рамках арифметики с плавающей точкой, сохраняя пять разрядов числа. После первого шага исключения получим

$$\begin{aligned} 10x_1 - 7x_2 &= 7, \\ -0.001x_2 + 6x_3 &= 6.001, \\ 2.5x_2 + 5x_3 &= 2.5. \end{aligned}$$

Следующий шаг исключения проводим при малом ведущем элементе (-0.001). Чтобы исключить x_2 из третьего уравнения, мы вынуждены умножить второе уравнение на 2500. При умножении получаем число 15002,5, которое нужно округлить до пяти разрядов. В результате получаем третье уравнение в виде

$$15005x_3 = 15006.$$

Отсюда x_3 = 15 006/15 005 = 1.0001. Из второго и первого уравнений найдем

$$x_2 = \frac{6.001 - 6 \cdot 1.0001}{-0.001} = -0.4, \quad x_1 = \frac{7 + 7 \cdot (-0.4)}{10} = 0.42.$$

Вычисления проводились с усечением до пяти разрядов по аналогии с процессом вычислений на компьютере. В результате этого было получено решение (0.42, - 0.4, 1.0001) вместо (0, - 1,1).

Такая большая неточность результатов объясняется малой величиной ведущего элемента. В подтверждение этому переставим сначала уравнения системы:

$$\begin{aligned} 10x_1 - 7x_2 &= 7, \\ 2.5x_2 + 5x_3 &= 2.5, \\ -0.001x_2 + 6x_3 &= 6.001. \end{aligned}$$

Исключим теперь x_2 из третьего уравнения, прибавив к нему второе, умноженное на 0.0004 (ведущий элемент

здесь равен 2.5). Третье уравнение примет вид

$$6.002x_2 = 6.002.$$

Отсюда находим $x_3 = 1$. С помощью второго и первого уравнений вычислим x_2, x_1:

$$x_2 = \frac{2.5 - 5\cdot 1}{2.5} = -1,\ \ x_1 = \frac{7 + 7\cdot(-1)}{10} = 0.$$

Таким образом, в результате перестановки уравнений, т. е. выбора наибольшего по модулю из оставшихся в данном столбце элементов, погрешность решения в рамках данной точности исчезла.

Рассмотрим подробнее вопрос о погрешностях решения систем линейных уравнений методом Гаусса. Запишем систему в матричном виде: $Ax = b$. Решение этой системы можно представить в виде $x=\boldsymbol{A}^{-1}b$. Однако вычисленное по методу Гаусса решение X^* отличается от этого решения из-за погрешностей округлений, связанных с ограниченностью разрядной сетки машины.

JСуществуют две величины, характеризующие степень отклонения полученного решения от точного. Одна из них — *погрешность* ΔX, равная разности этих значений, другая — *невязка* τ, равная разности между правой и левой частями уравнений при подстановке в них решения:

$$\Delta x = x - x_*, \quad \tau = Ax_* - b.$$

Можно показать, что если одна из этих величин равна нулю, то и другая должна равняться нулю. Однако из малости одной не следует малость другой. При $\Delta X \approx 0$ обычно $\tau \approx 0$, но обратное утверждение справедливо не всегда. В частности, для плохо обусловленных систем при $\tau \approx 0$ погрешность решения может быть большой.

Вместе с тем в практических расчетах, если система не является плохо обусловленной, контроль точности решения осуществляется с помощью невязки. Можно отметить, что метод Гаусса с выбором главного элемента в этих случаях дает малые невязки.

1.3 Определитель и обратная матрица.

Ранее уже отмечалось, что непосредственное нахождение определителя требует большого объема вычислений. Вместе с тем легко вычисляется определитель треугольной матрицы: он равен произведению ее диагональных элементов.

Для приведения матрицы к треугольному виду может быть использован *метод исключения,* т. е. прямой ход метода Гаусса. В процессе исключения элементов величина определителя не меняется. Знак определителя меняется на противоположный при перестановке его столбцов или строк. Следовательно, значение определителя после приведения матрицы A к треугольному виду вычисляется по формуле

$$\det A = (-1)^k \prod_{i=1}^{n} a_{ii}.$$

Здесь диагональные элементы a_{ii} берутся из преобразованной (а не исходной) матрицы. Знак зависит от того, четной или нечетной была суммарная перестановка строк (или столбцов) матрицы при ее приведении к треугольному виду (для получения ненулевого или максимального по модулю ведущего элемента на каждом этапе исключения). Благодаря методу исключения можно вычислять определители 1000-го и большего порядков, и объем вычислений значительно меньший, чем в проведенных ранее оценках.

Теперь найдем обратную матрицу $\boldsymbol{A}^{-1}$. Обозначим ее элементы через z_{ij}. Запишем равенство $\boldsymbol{A}\boldsymbol{A}^{-1} = E$ в виде

$$\begin{pmatrix} a_{11} & a_{12} \ldots a_{1n} \\ a_{21} & a_{22} \ldots a_{2n} \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ a_{n1} & a_{n2} \ldots a_{nn} \end{pmatrix} \cdot \begin{pmatrix} z_{11} & z_{12} \ldots z_{1n} \\ z_{21} & z_{22} \ldots z_{2n} \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ z_{n1} & z_{n2} \ldots z_{nn} \end{pmatrix} = \begin{pmatrix} 1\ 0\ \ldots\ 0 \\ 0\ 1\ \ldots\ 0 \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ 0\ 0\ \ldots\ 1 \end{pmatrix}. \qquad (4.12)$$

Отсюда следует, что

$$Az_i = e_j, \quad j = 1, 2, \ldots, n,$$

Где z_1 и e_i-j-е столбцы матриц $\boldsymbol{A}^{-1}$ и E соответственно. Таким образом, для нахождения j-го столбца обратной матрицы нужно решить систему уравнений (5.12). Решив n таких систем для j=1,2,…, n, мы найдём все столбцы z_i и, следовательно, саму обратную матрицу.

Поскольку при разных j матрица A системы (5.12) не меняется, исключение неизвестныхпри использовании метода Гаусса (прямой ход) проводится только один раз, причем сразу для всех правых частей – столбцов e_i. Затем для каждой из систем (5.12) делается обратный ход в соответствующей преобразованной правой частью.

Оценки показывают, что это весьма экономичный способ обращения матрицы. Он требует примерно лишь в три раза больше действий, чем при решении одной системы уравнений.

Степень отклонения вычисленной обратной матрицы $\boldsymbol{A}_*^{-1}$ от её точного значения характеризуется погрешностью $\Delta\boldsymbol{A}^{-1}$ и невязкой R:

$$\Delta A^{-1} = A^{-1} - A_*^{-1}, \ \ R = AA_*^{-1} - E.$$

1.4 Метод прогонки

Он является модификацией метода Гаусса для частного случая разреженных систем — системы уравнений с трехдиагональной матрицей. Такие системы получаются при моделировании некоторых инженерных задач, а также при численном решении краевых задач для дифференциальных уравнений.

Запишем систему уравнений в виде

$$\begin{aligned}
&b_1x_1 + c_1x_2 && = d_1,\\
&a_2x_1 + b_2x_2 + c_2x_3 && = d_2,\\
&\quad a_3x_2 + b_3x_3 + c_3x_4 && = d_3,\\
&\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\\
&\qquad a_{n-1}x_{n-2} + b_{n-1}x_{n-1} + c_{n-1}x_n && = d_{n-1},\\
&\qquad\qquad a_nx_{n-1} + b_nx_n && = d_n.
\end{aligned} \tag{4.13}$$

На главной диагонали матрицы этой системы стоят элементы b_1, b_2, ..., b_n, над ней — элементы c_1, c_2, ..., c_{n-1} под ней — элементы a_2, a_3, ..., a_n. (при этом обычно все коэффициенты b_i не равны нулю.

Метод прогонки состоит из двух этапов — *прямой прогонки* (аналога прямого хода метода Гаусса) и *обратной прогонки* (аналога обратного хода метода Гаусса). Прямая прогонка состоит в вычислении прогоночных коэффициентов A_i,B_i, с помощью которых каждое неизвестное x_i выражается через x_{i+1}:

$$x_i = A_ix_{i+1} + B_i,\ \ i = 1, 2, ..., n-1. \tag{4.14}$$

Из первого уравнения системы (5.13) найдем

$$x_1 = -\frac{c_1}{b_1}x_2 + \frac{d_1}{b_1}. \tag{4.15}$$

С другой стороны, по формуле (5.14) $x_1=A_1x_2 + B_1$. Приравнивая коэффициенты в обоих выражениях для x_1 получаем

$$A_1 = -\frac{c_1}{b_1},\ \ B_1 = \frac{d_1}{b_1}.$$

Подставим во второе уравнение системы (5.13) вместо x_1 его выражение через x_2 по формуле (5.14)

$$a_2(A_1x_2 + B_1) + b_2x_2 + c_2x_3 = d_2.$$

Выразим отсюда x_2 через x_3:

$$x_2 = \frac{-c_2x_3 + d_2 - a_2B_1}{a_2A_1 + b_2}, \tag{4.16}$$

или

$$x_2 = A_2x_3 + B_2,$$

$$A_2 = -\frac{c_2}{e_2},\ \ B_2 = \frac{d_2 - a_2B_1}{e_2},\ \ e_2 = a_2A_1 + b_2.$$

Аналогично можно вычислить прогоночные коэффициенты для любого номера i:

$$A_i = -\frac{c_i}{e_i},\ \ B_i = \frac{d_i - a_iB_{n-1}}{e_i},$$

$$e_i = a_iA_{i-1} + b_i,\ \ i = 2, 3, ..., n-1.$$

Обратная прогонка состоит в последовательном вычислении неизвестных x_i. Сначала нужно найти x_n. Для этого воспользуемся выражением (4.14) при i=n-1 и последним уравнением системы (4.13). Запишем их:

$$x_{n-1} = A_{n-1}x_n + B_{n-1},$$

$$a_nx_{n-1} + b_nx_n = d_n.$$

Отсюда, исключая x_{n-1}, находим

$$x_n = \frac{d_n - a_nB_{n-1}}{b_n + a_nA_{n-1}}.$$

Далее, используя формулы (5.14) и вычисленные ранее по формулам (5.15), (5.16) прогоночные коэффициенты, последовательно вычисляем все неизвестные x_{n-1}, x_{n-2}, …, x_1 Алгоритм решения системы линейных уравнений вида (5.13) методом прогонки представлен на рисунке 5.3.

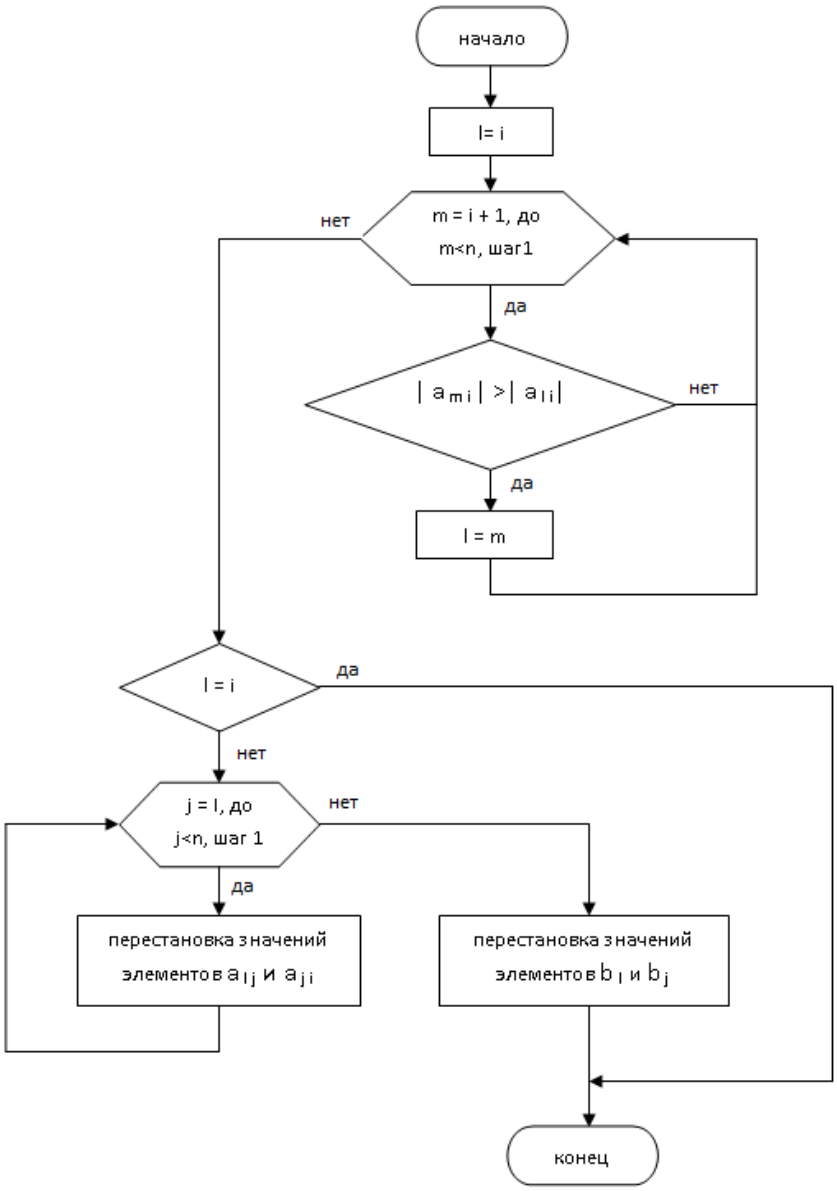

Рисунок 4.3

При анализе алгоритма метода прогонки надо учитывать возможность деления на нуль в формулах (5.15), (5.16). Можно показать, что при выполнении условия преобладания диагональных элементов, т. е. если $|b_i| \geq |a_i| + |c_i|$, причем хотя бы для одного значения i имеет место строгое неравенство, деления на нуль не возникает, и система (5.13) имеет единственное решение.

Приведенное условие преобладания диагональных элементов обеспечивает также устойчивость метода прогонки относительно погрешностей округлений. Последнее обстоятельство позволяет использовать метод прогонки для решения больших систем уравнений. Заметим, что данное условие устойчивости прогонки является достаточным, но не необходимым. В ряде случаев для хорошо обусловленных систем вида (5.13) метод прогонки оказывается устойчивым даже при нарушении условия преобладания диагональных элементов.

§2. Итерационные методы. Уточнение решения. Метод простой итерации. Метод Гаусса-Зейделя

2.1 Уточнение решения.

Решения, получаемые с помощью прямых методов, обычно содержат погрешности, вызванные округлениями при выполнении операций над числами с плавающей точкой на компьютере с ограниченным числом разрядов. В ряде случаев эти погрешности могут быть значительными, и необходимо найти способ их уменьшения. Рассмотрим здесь один из методов, позволяющий уточнить решение, полученное с помощью прямого метода.

Найдем решение системы линейных уравнений

$$Ax=b \tag{4.17}$$

Пусть с помощью некоторого прямого метода вычислено приближенное решение $x^{(0)}$ (т. е. приближенные значения неизвестных $x_1^{(0)}$, $x_2^{(0)}$,…, $x_n^{(0)}$), называемое начальным или нулевым приближением к решению. Подставляя это решение в левую часть системы (4.17) получаем некоторый столбец правых частей $b^{(0)}$, отличный от b:

$$Ax^{(0)}=b^{(0)} \tag{4.18}$$

Введем обозначения: $\Delta x^{(0)}$— погрешности значений неизвестных, $\tau^{(0)}$ невязка, т. е.

$$\Delta x^{(0)} = x - x^{(0)}, \ \tau^{(0)} = Ax^{(0)} - b = b^{(0)} - b. \tag{4.19}$$

Вычитая равенство (4.18) из равенства (5.17), с учетом обозначений (4.19) получаем

$$A\Delta x^{(0)} = -\tau^{(0)} \tag{4.20}$$

Решая эту систему, находим значения погрешностей: $\Delta x^{(0)}$,; которое используем в качестве поправки к приближенному решению $x^{(0)}$, вычисляя таким образом новое приближенное решение $x^{(1)}$ (или следующее приближение к решению):

$$x^{(1)}=x^{(0)}+\Delta x^{(0)}.$$

Таким же способом можно найти новую поправку к решению $\Delta x^{(1)}$ и следующее приближение $x^{(2)}=x^{(1)}+\Delta x^{(1)}$ и т. д. Процесс продолжается до тех пор, пока все очередное значение погрешности (поправки) $\Delta x^{(k)}$ не станет достаточно малым, т. е. пока очередные приближенные значения неизвестных $x_1^{(k+1)}, x_2^{(k+1)}, \dots, x_n^{(k+1)}$ не будет мало отличаться от предыдущих значений $x_1^{(k)}, x_2^{(k)}, \dots, x_n^{(k)}$.

Рассмотренный процесс уточнения решения представляет собой фактически итерационный метод решения системы линейных уравнений. При этом заметим, что для нахождения очередного приближения, т.е. на каждой итерации, решаются системы уравнений вида (5.20) с одной и той же матрицей, являющейся матрицей исходной системы (5.17), при разных правых частях. Это позволяет строить экономичные алгоритмы. Например, при использовании метода Гаусса сокращается объем вычислений на этапе прямого хода.

Решение систем линейных уравнений с помощью рассмотренного метода (а также решение систем линейных уравнений иными итерационными методами, решение итерационными методами уравнений другого вида и их систем) сводится к следующему (Рисунок 4.4). Вводятся исходные данные, например, коэффициенты уравнений и допустимое значение погрешности. Необходимо также задать начальные приближения значений неизвестных (вектор-столбец $x^{(0)}$). Они либо водятся в компьютер, либо вычисляются каким-либо способом (в частности, путем решения системы уравнений с помощью прямого метода). Затем организуется циклический вычислительный процесс, каждый цикл которого представляет собой одну итерацию — переход от предыдущего приближениях $x^{(k-1)}$ к последующему $x^{(k)}$. Если оказывается, что с увеличением числа итераций приближенное решение стремится к точному:

$$\lim_{k\to\infty}^{(k)} = x,$$

то итерационный метод называют *сходящимся.*

На практике наличие сходимости и достижение требуемой точности обычно определяют приближенно, поступая следующим образом. При малом (с заданной допустимой погрешностью) изменении x на двух последовательных итерациях, т. е. при малом отличии $x^{(k)}$ от $x^{(k-1)}$, процесс прекращается, и происходит вывод значений неизвестных, полученных на последней итерации.

Возможны разные подходы к определению малости отличия x на двух последовательных итерациях.

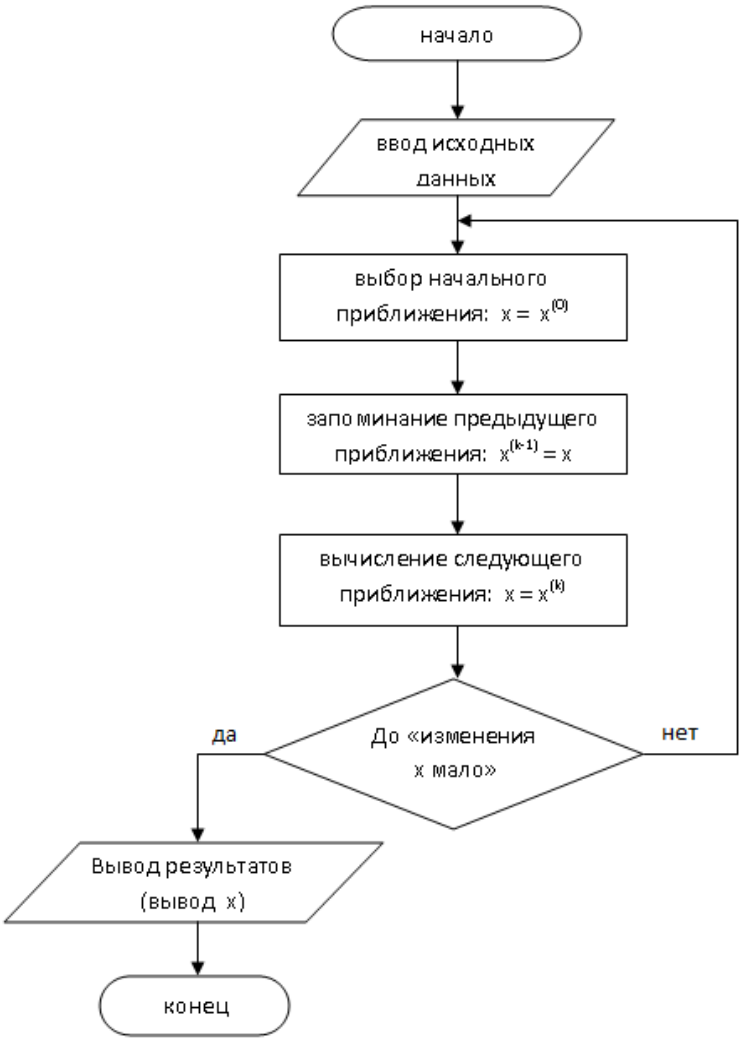

Рисунок 4.4

Например, если задана допустимая погрешность $\varepsilon > 0$, то критерием окончания итерационного процесса можно считать выполнение одного из трех неравенств:

$$| x^{(k)} - x^{(k-1)} | = \sqrt{\sum_{i=1}^{n} \left(x_i^{(k)} - x_i^{(k-1)} \right)^2} < \varepsilon, \tag{4.21}$$

$$\max_{1 \le i \le n} | x_i^{(k)} - x_i^{(k-1)} | < \varepsilon, \tag{4.22}$$

$$\max_{1 \le i \le n} \left| \frac{x_i^{(k)} - x_i^{(k-1)}}{x_i^{(k)}} \right| < \varepsilon, \ при \ | x_i | >> 1. \tag{4.23}$$

Здесь в первом случае отличие векторов $x^{(k)}$ и $x^{(k-1)}$ «на ε» понимался в смысле малости модуля их разности, во втором - в смысле малости разностей всех соответствующих компонент векторов, в третьем — в смысле малости относительных разностей компонент. Если система не является плохо обусловленной, то в качестве критерия окончания итерационного процесса можно использовать и условие малости невязки, например

$$| r^{(k)} | < \varepsilon. \tag{4.24}$$

Заметим, что в рассмотренном алгоритме не предусмотрен случай отсутствия сходимости. Для предотвращения непроизводительных затрат машинного времени в алгоритм вводят счетчик числа итераций и при достижении им некоторого заданного значения счет прекращают. Такой элемент будет в дальнейшем введен в структурограмму.

2.2 Метод простой итерации.

Этот метод широко используется для численного решения уравнений и их систем различных видов. Рассмотрим применение метода простой итерации к решению систем линейных уравнений.

Запишем исходную систему уравнений в векторно-матричном виде, выполним ряд тождественных преобразований:

$$\begin{aligned} &Ax = b; \quad 0 = b - Ax; \quad x = b - Ax + x; \\ &x = (b - Ax)\tau + x; \quad x = (E - \tau A)x + \tau b; \\ &x = Bx + \tau b, \end{aligned} \tag{4.25}$$

где $\tau \neq 0$ – некоторое число, Е — единичная матрица, $B=E-\tau A$. получившаяся система (5.25) эквивалентна исходной системе и служит основой для построения метода простой итерации.

Выберем некоторое начальное приближение $x^{(0)}$ и подставим его в правую часть системы (5.25):

$$x^{(1)} = Bx^{(0)} + \tau b.$$

Поскольку $x^{(0)}$не является решением системы, в левой части (5.25) получится некоторый столбец $x^{(1)}$, в общем случае отличный от $x^{(0)}$. Полученный столбец $x^{(1)}$, будем рассматривать в качестве следующего (первого) приближения к решению. Аналогично, по известному k - му приближению можно найти (k+1)-е приближение:

$$x^{(k+1)} = Bx^{(x)} + \tau b, \;\; k = 0,1,2,... \tag{4.26}$$

Формула (4.26) и выражает собой метод простой итерации. Для ее применения нужно задать неопределенный пока параметр τ. От значения τ зависит, будет ли сходиться метод, а если будет, то какова будет *скорость сходимости* т. е. как много итераций нужно совершить для достижения требуемой точности. В частности, справедлива следующая теорема.

Теорема. Пусть det A$\neq$0. Метод простой итерации (4.26) сходится тогда и только тогда, когда все собственные числа матрицы B=A-τE по модулю меньше единицы.

Для некоторых типов матрицы A можно указать правило выбора τ, обеспечивающее сходимость метода и оптимальную скорость сходимости. В простейшем же случае τ можно положить равным некоторому постоянному числу, например, 1, 0.1 и т. д.

2.3 Метод Гаусса-Зейделя.

Одним из самых распространенных итерационных методов, отличающийся простотой и легкостью программирования, является *метод Гаусса-Зейделя.*

Проиллюстрируем сначала этот метод па примере решения системы

$$\begin{aligned} &a_{11}x_1 + a_{12}x_1 + a_{13}x_3 = b_1, \\ &a_{21}x_1 + a_{22}x_1 + a_{23}x_3 = b_2, \\ &a_{31}x_1 + a_{32}x_1 + a_{33}x_3 = b_3. \end{aligned} \tag{4.27}$$

Предположим, что диагональные элементы *ац, a_{22}, озз* отличны от нуля (в противном случае можно переставить уравнения). Выразим неизвестные x_1, x_2, и x_3 в соответственно из первого, второго и третьего уравнений системы (4.27):

$$x_1 = \frac{1}{a_{11}}\left(b_1 - a_{12}x_2 - a_{13}x_3\right), \tag{4.28}$$

$$x_2 = \frac{1}{a_{22}}\left(b_2 - a_{21}x_1 - a_{23}x_3\right), \tag{4.29}$$

$$x_3 = \frac{1}{a_{33}}\left(b_3 - a_{31}x_1 - a_{32}x_2\right), \tag{4.30}$$

Зададим некоторые начальные (нулевые) приближения значений неизестных: $x_1 = x_1^{(0)}, x_2 = x_2^{(0)}, x_3 = x_3^{(0)}$. Подставляя эти значения в правую часть выражения (4.28), получаем новое (первое) приближение для x_1:

$$x_1^{(1)} = \frac{1}{a_{11}}\left(b_1 - a_{12}x_2^{(0)} - a_{13}x_3^{(0)}\right).$$

Используя это значение для x_1 и приближение $x_3^{(0)}$ для x_3, находим из (5.29) первое приближение для x_2:

$$x_2^{(1)} = \frac{1}{a_{22}}\left(b_2 - a_{21}x_1^{(1)} - a_{23}x_3^{(0)}\right).$$

И наконец, используя вычисленные значения $x_1 = x_1^{(0)}, x_2 = x_2^{(0)},$ находим с помощью выражения (4.30) первое приближение для x_3:

$$x_3^{(1)} = \frac{1}{a_{33}}\left(b_3 - a_{31}x_1^{(1)} - a_{32}x_2^{(1)}\right).$$

На этом заканчивается первая итерация решения системы (5.28) - (5.30). Используя теперь значения $x_1^{(1)}$, $x_2^{(1)}, x_3^{(1)}$, можно таким же способом провести вторую итерацию, в результате которой будут найдены вторые приближения к решению: $x_1 = x_1^{(2)}, x_2 = x_2^{(2)}, x_3 = x_3^{(2)}$ и т. д.

Приближение с номером *k* можно вычислить, зная приближение с номером *k-1*, как

$$x_1^{(k)} = \frac{1}{a_{11}}\left(b_1 - a_{12}x_2^{(k-1)} - a_{13}x_3^{(k-1)}\right),$$

$$x_2^{(k)} = \frac{1}{a_{22}}\left(b_2 - a_{21}x_1^{(k)} - a_{23}x_3^{(k-1)}\right),$$

$$x_3^{(k)} = \frac{1}{a_{33}}\left(b_3 - a_{31}x_1^{(k)} - a_{32}x_2^{(k)}\right).$$

Итерационный процесс продолжается до тех пор, пока значения $x_1^{(k)}, x_2^{(k)}, x_3^{(k)}$ не станут близкими с заданной погрешностью к значениям $x_1^{(k-1)}, x_2^{(k-1)}, x_3^{(k-1)}$.

ПРИМЕР. Решить с помощью метода Гаусса-Зейделя следующую систему уравнений:

$$4x_1 - x_2 + x_3 = 4,$$
$$2x_1 + 6x_2 - x_3 = 7,$$
$$x_1 + 2x_2 - 3x_3 = 0.$$

Легко проверить, что решение данной системы следующее: $x_1 = x_2 = x_3 = 1.$

Решение. Выразим неизвестные x_1, x_2 и x_3 соответственно из первого, второго и третьего уравнений:

$$x_1 = \frac{1}{4}\left(4 + x_2 - x_3\right), \quad x_2 = \frac{1}{6}\left(7 - 2x_1 + x_3\right),$$

$$x_3 = \frac{1}{3}(x_1 + 2x_2).$$

В качестве начального приближения (как это обычно делается) примем $x_1^{(0)} = 0,\ x_2^{(0)} = 0,\ x_3^{(0)} = 0.$ Найдем новые приближения неизвестных:

$$x_1^{(1)} = \frac{1}{4}(4 + 0 - 0) = 1, \quad x_2^{(1)} = \frac{1}{6}(7 - 2\cdot 1 + 0) = \frac{5}{6},$$

$$x_3^{(1)} = \frac{1}{3}(1 + 2\cdot\frac{5}{6}) = \frac{8}{9}.$$

Аналогично вычислим следующие приближения:

$$x_1^{(2)} = \frac{1}{4}(4 + \frac{5}{6} - \frac{8}{9}) = \frac{71}{72}, \quad x_2^{(2)} = \frac{1}{6}(7 - 2\cdot\frac{71}{72} + \frac{8}{9}) = \frac{71}{72},$$

$$x_3^{(2)} = \frac{1}{3}(\frac{71}{72} + 2\cdot\frac{71}{72}) = \frac{71}{72}.$$

Итерационный процесс можно продолжать до получения малой разности между значениями неизвестных в двух последовательных итерациях. Рассмотрим теперь систему *n* линейных уравнений с n неизвестными. Запишем ее в виде

$$a_{i1}x_1 + \ldots + a_{i,i-1}x_{i-1} + a_{ii}x_i + a_{i,i+1}x_{i+1} + \ldots + a_{in}x_n = b_i,$$
$$i = 1, 2, \ldots, n.$$

Здесь также будем, предполагать, что все диагональные элементы отличны от нуля. Тогда в соответствии с методом Гаусса-Зейделя *k-с* приближение к решению можно представить в виде

$$x_i^{(k)} = \frac{1}{a_{ii}}(b_i - a_{i1}x_1^{(k)} - \ldots - a_{i,i-1}x_{i-1}^{(k)} - a_{i,i+1}x_{i+1}^{(k-1)} - \ldots - a_{in}x_n^{(k-1)}),$$
$$i = 1, 2, \ldots, n. \qquad (4.31)$$

Итерационный процесс продолжается до тех пор, пока все значения $x_i^{(k)}$ не станут близкими к $x_i^{(k-1)}$, т. е. в качестве критерия завершения итераций используется одно из условий (5.21) - (5.23), (5.24).

Для сходимости итерационного процесса (5.31) достаточно, чтобы модули диагональных коэффициентов для каждого уравнения системы были не меньше сумм модулей всех остальных коэффициентов (преобладание диагональных элементов):

$$|a_{ii}| \geq \sum_{j \neq i} |aij|, \; i = 1, 2, \ldots, n. \qquad (4.32)$$

При этом хотя бы для одного уравнения неравенство должно выполняться строго. Эти условия являются достаточными для сходимости метода, но они не являются необходимыми, т. е. для некоторых систем итерации сходятся и при нарушении условий (4.32).

Алгоритм решения системы *n* линейных уравнений методом Гаусса-Зейделя представлен на рисунке 5.5. В качестве исходных данных вводятся n, коэффициенты и правые части уравнений системы, погрешность ε, максимально допустимое число итераций M, а также начальные приближения переменных x_i (i = 1,2,...,n). Отметим, что начальные приближения можно не вводить в компьютер, а полагать их равными некоторым значениям (например, нулю). Критерием завершения итераций выбрано условие (22),

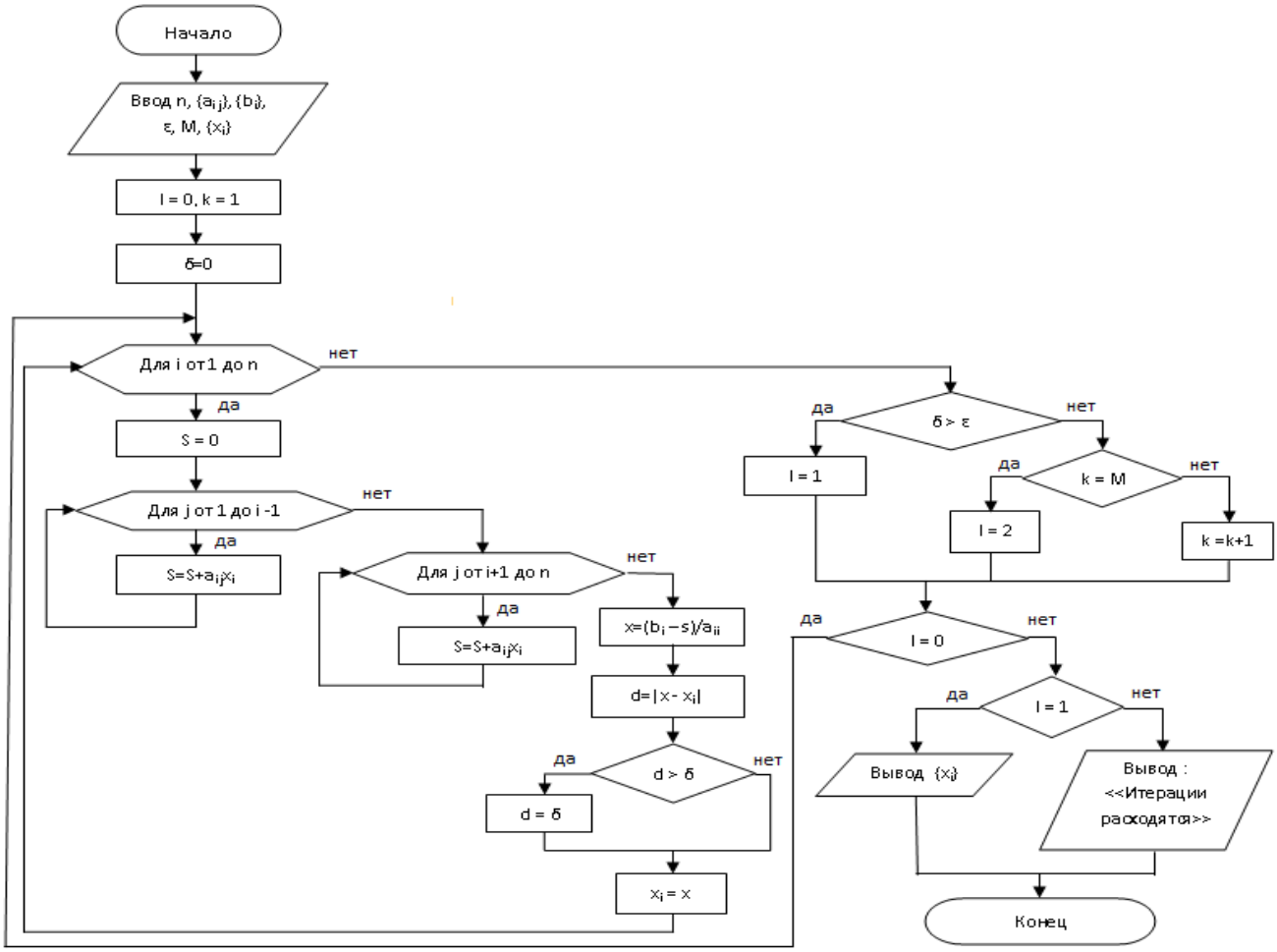

Рисунок 4.5

в котором через δ обозначена максимальная абсолютная величина разности $x_i^{(k)}$ и $x_i^{(k-1)}$:

$$\delta = \max_{1 \leq i \leq n} |x_i^{(k)} - x_i^{(k-1)}| < \varepsilon$$

Для удобства чтения струкгурограммы объясним другие обозначения: *k*- порядковый номер итерации; *i* — номер уравнения, а также переменного, которое вычисляется в соответствующем цикле; *j* — номер члена вида в правой части соотношения (5.31). Итерационный процесс прекращается либо при $\delta < \varepsilon$, либо при $\kappa = M$. В последнем случае итерации не сходятся, о чем выдается сообщение. Для завершения цикла, реализующего

итерационный процесс, используется переменная *l*, которая принимает значения 0, 1 и 2 соответственно при продолжении итераций, при выполнении условия $\delta < \varepsilon$ и при выполнении условия *k=M.*

§3. Задачи на собственные значения. Метод вращений. Трехдиагональные матрицы.

3.1 Задачи на собственные значения.

Основные понятия. Большое число научно-технических задач, а также некоторые исследования в области вычислительной математики требуют нахождения собственных значений и собственных векторов матриц. Введем некоторые определения, необходимые для изложения материала данного параграфа.

Рассмотрим, квадратную матрицу n-го

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}. \tag{4.33}$$

Вектор x=$\{x_1, x_2, \dots, x_n\}$ называется *собственным вектором* матрицы A соответствующим *собственному значению* λ, если он удовлетворяет системе уравнений

$$Ax = \lambda x. \tag{4.34}$$

Поскольку при умножении собственного вектора на скаляр он остается собственным вектором той же матрицы, его можно нормировать. В частности, каждую координату собственного вектора можно разделить на максимальную из них или на длину вектора; в последнем случае получится единичный собственный вектор.

Характеристической матрицей C данной матрицы *A* называется матрица вида

$$C = A - \lambda E = \begin{pmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} - \lambda \end{pmatrix} \tag{4.35}$$

где *E* — единичная матрица. Легко видеть, что систему (5.34) можно записать в виде

$$(A - \lambda E)x = 0 \;\; или \;\; Cx = 0. \tag{4.36}$$

Если перейти к координатной форме записи вектора x, то с учетом (5.33) систему (5.36) можно записать в виде

$$\begin{aligned} &(a_{11} - \lambda)x_1 + a_{12}x_2 + \dots + a_{1n}x_n = 0, \\ &a_{21}x_1 + (a_{22} - \lambda)x_2 + \dots + a_{2n}x_n = 0, \\ &\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots, \\ &a_{n1}x_1 + a_{n2}x_2 + \dots + (a_{nn} - \lambda)x_n = 0. \end{aligned} \tag{4.37}$$

Система (4.36) или (5.37) является однородной системой *n* линейных уравнений с *n* неизвестными. Она имеет ненулевые решения лишь тогда, когда ее определитель равен нулю: det *C* =0, причем решение не единственно. Определитель матрицы *C* является многочленом n-й степени относительно λ:

$$\det C = c_0\lambda^n + c_1\lambda^{n-1} + \dots + c_{n-1}\lambda + c_n, \tag{4.38}$$

называемым *характеристическим многочленом.* Корни этого многочлена являются собственными значениями матрицы *A*.

Для нахождения собственных векторов матрицы требуется решить систему линейных алгебраических уравнений, решение которой не единственно. Из линейной алгебры известно, что в этом случае структура общего решения системы имеет следующий вид: одно или несколько неизвестных, называемых *свободными,* могут принимать любые значения, а остальные неизвестные выражаются через свободные. Число свободных неизвестных равно числу уравнений системы, являющихся следствием остальных уравнений. На практике, если свободное неизвестное одно (что часто и бывает), его полагают равным некоторому числу, например единице. После этого остальные неизвестные (компоненты вектора) находятся однозначно из подсистемы линейно независимых уравнений, в которой отброшено уравнение, являющееся следствием остальных. Эта процедура не влияет на результат решения задачи, поскольку, как уже отмечалось, собственные векторы находятся с точностью до постоянного множителя.

Пример. Вычислить собственные числа и собственные векторы матрицы

$$A=\begin{pmatrix}3 & 1\\ 2 & 4\end{pmatrix}.$$

Решение. Составим характеристический многочлен

$$\begin{vmatrix}3-\lambda & 1\\ 2 & 4-\lambda\end{vmatrix}=(3-\lambda)(4-\lambda)-2=\lambda^2-7\lambda+10.$$

Найдем корни этого многочлена второй степени:

$$\lambda^2-7\lambda+10=0,\ \lambda_1=2,\ \lambda_2=5.$$

Дня нахождения собственных векторов x_1, x_2, соответствующих собственным значениям λ_1, λ_2 составим системы уравнений типа (5.36), (5.37) для каждого из них.

При λ_1 =2 получим

$$\begin{pmatrix}3-2 & 1\\ 2 & 4-2\end{pmatrix}\cdot\begin{pmatrix}x_1\\ x_2\end{pmatrix}=\begin{pmatrix}0\\ 0\end{pmatrix},$$

или, в координатной форме,

$$x_1+x_2=0,$$
$$2x_1+2x_2=0.$$

Замечаем, что уравнения линейно зависимы. Поэтому оставляем лишь одно из них.

Из первого уравнения следует, что $x_2=-x_1$. Неизвестное x_1 можно считать свободным, полагаем $x_1=1$. Тогда $x_2=-1$, и собственный вектор, соответствующий собственному значению λ_1 =2, имеет вид $x_1=\{1, -1\}$ или $x_1=e_1-e_2$, где e_1, e_2 единичные орты выбранной базисной системы.

Аналогично находим второй собственный вектор, соответствующий собственному значению λ_2=5. Опуская комментарии, получаем

$$\begin{pmatrix}3-5 & 1\\ 2 & 4-5\end{pmatrix}\cdot\begin{pmatrix}x_1\\ x_2\end{pmatrix}=\begin{pmatrix}0\\ 0\end{pmatrix},\quad \begin{matrix}-2x_1+x_2=0,\\ 2x_1-x_2=0.\end{matrix}$$

Отсюда $x_1=1$, $x_2=2$, $x_2=e_1+2e_2$.

Вектор x_1 нормирован; нормируем также вектор x_2, разделив его компоненты на наибольшую из них. Получим $x_2=0.5e_1+e_2$. Можно также привести векторы к единичной длине, разделив их компоненты на значения модулей векторов. В этом случае

$$x_1=\frac{1}{\sqrt{2}}(e_1-e_2),\ x_2=\frac{1}{\sqrt{5}}(e_1+2e_2).$$

Мы рассмотрели простейший пример вычисления собственных значений и собственных векторов для матрицы второго порядка. Нетрудно также провести подобное решение задачи для матрицы третьего порядка и для некоторых весьма специальных случаев.

В общем случае, особенно для матриц высокого порядка, задача о нахождении их собственных значений и собственных векторов, называемая *полной проблемой собственных значений,* значительно более сложная.

На первый взгляд может показаться, что вопрос сводится к вычислению корней многочлена (5.38). Однако здесь задача осложнена тем, что среди собственных значений часто встречаются кратные. И кроме того, для произвольной матрицы непросто вычислить сами коэффициенты характеристического многочлена.

Отметим некоторые *свойства собственных значений* для частных типов исходной матрицы.

1. Все собственные значения симметрической матрицы действительны.
2. Если собственные значения матрицы действительны и различны, то соответствующие им собственные векторы ортогональны и образуют базис рассматриваемого пространства. Следовательно, любой вектор в данном пространстве можно выразить через совокупность линейно независимых собственных векторов.
3. Если две матрицы *A* и *B подобны,* т. е. они связаны соотношением

$$B = P^{-1}AP, \tag{4.39}$$

то их собственные значения совпадают (здесь P — некоторая матрица). *Преобразование подобия* (5.39) можно использовать для упрощения исходной матрицы, а задачу о вычислении ее собственных значений свести к аналогичной задаче для более простой матрицы. Очевидно, самым лучшим упрощением матрицы (5.33) было бы приведение ее к треугольному виду

$$A' = \begin{pmatrix} a'_{11} & a'_{12} & \dots & a'_{1n} \\ & a'_{22} & \dots & a'_{2n} \\ & & \dots & \dots \\ 0 & & & a'_{nn} \end{pmatrix}.$$

Тогда матрица (5.35) также имела бы треугольный вид. Как известно, определитель треугольной матрицы равен произведению ее диагональных элементов, поэтому характеристический многочлен (5.38) в этом случае имеет вид

$$\det C = (a'_{11} - \lambda)(a'_{22} - \lambda)\dots(a'_{nn} - \lambda).$$

Собственные значения матрицы, равные корням этого многочлена, можно получить сразу:

$$\lambda_1 = a'_{11},\ \lambda_2 = a'_{22},\ \dots,\ \lambda_n = a'_{nn}.$$

Таким образом, собственные значения треугольной матрицы равны ее диагональным элементам. То же самое, естественно, относится и к диагональной матрице, которая является частным случаем треугольной.

Некоторые типы матриц удается привести к треугольному виду с помощью преобразования подобия. В частности, симметрическую матрицу можно привести к диагональному виду. На практике часто используется приведение симметрической матрицы к трехдиагональному виду. Процедура вычисления собственных значений для полученной матрицы значительно упрощается по сравнению с задачей для исходной матрицы.

Существует ряд методов, основанных на использовании преобразования подобия, позволяющего привести исходную матрицу к более простой структуре. Мы рассмотрим ниже один из них — метод вращений.

3.2 Метод вращений.

Одним из эффективных методов, позволяющих привести исходную симметрическую матрицу n-го порядка к трехдиагональному виду, является *метод вращений.* Он основан на специально подбираемом вращении системы координат в n-мерном пространстве. Поскольку любое вращение можно заменить последовательностью элементарных (плоских) вращений, то решение задачи можно разбить на ряд шагов, на каждом из которых осуществляется плоское вращение. Таким образом, на каждом шаге выбираются две оси — *i*-я и j-я (i≠j), и поворот на угол φ производится в плоскости, проходящей через эти оси; остальные оси координат на данном шаге неподвижны. *Матрица вращения* мри этом имеет вид

$$P_{ij} = \begin{pmatrix} 1 & & & & & & \\ & 1 & & & & 0 & \\ & & \ddots & & & & \\ & & & p_{ii} & \cdots & p_{ij} & \\ & & & \vdots & \ddots & \vdots & \\ & & & p_{ji} & \cdots & p_{jj} & \\ & & & & & & \ddots \\ & 0 & & & & 1 & \\ & & & & & & 1 \end{pmatrix}, \tag{4.40}$$

$$p_{ii} = p_{jj} = p,\ \ p_{ij} = -p_{ji} = -q,\ \ p = \cos\varphi,\ \ q = \sin\varphi.$$

Здесь мы рассматриваем матрицы с вещественными элементами. В случае комплексных векторов для использования этого метода нужно изменить формулы (5.40).

Для осуществления преобразования подобия (5.39) необходимо найти обратную матрицу P_{ij}^{-1}. Можно показать, что она равна в рассматриваемом случае транспонированной матрице P_{ij}^{T}. Для получения обратной матрицы достаточно провести зеркальное отражение всех элементов исходной матрицы относительно ее диагонали. Другими словами, нужно поменять местами строки и столбцы исходной матрицы; элементы p_{ij} и p_{ji} при этом поменяются местами.

Угол поворота φ на каждом шаге выбирается таким, чтобы в преобразованной матрице обратился в нуль один элемент (в симметрической матрице — два). Процесс преобразования исходной матрицы путем элементарного вращения на любом k-м шаге можно представить в виде рекуррентных соотношений

$$A^{(k)} = P_{ij}^T A^{(k-1)} P_{ij},\ k = 1, 2, \ldots \tag{4.41}$$

Рассмотрим первый шаг преобразования. Сначала вычисляется произведение матриц $B = A^{(0)} P_{ij}$ (здесь $A^{(0)}$- исходная матрица A). Как видно из (5.40), в полученной матрице отличными от исходных являются элементы, стоящие в *i*-м и *j*-м столбцах; остальные элементы совпадают с элементами матрицы $A^{(0)}$, т.е.

$$\begin{aligned} b_{li} &= a_{li}^{(0)} p + a_{lj}^{(0)} q,\ b_{lj} = -a_{li}^{(0)} q + a_{lj}^{(0)} p, \\ b_{lm=} a_{lm}^{(0)},\ &m \neq i, j,\ l = 1, 2, \ldots, n. \end{aligned} \tag{4.42}$$

Затем находится преобразованная матрица $A^{(1)} = P_{ij}^T B$. Элементы полученной матрицы отличаются от элементов матрицы *B* только *i-й* и *j-и* строками. Они связаны соотношениями

$$\begin{aligned} a_{im}^{(1)} &= b_{im} p + b_{jm} q,\ a_{jm}^{(1)} = -b_{im} q + b_{jm} p, \\ a_{lm}^{(1)} &= b_{lm},\ l \neq i, j,\ m = 1, 2, \ldots, n. \end{aligned} \tag{4.43}$$

Таким образом, преобразованная матрица $A^{(1)}$отличается от $A^{(0)}$ элементами строк и столбцов с номерами *i* и *j*. Эти элементы пересчитываются по формулам (4.42), (4.43). В данных формулах пока не определенными остались параметры: *p* и *q;* при этом лишь один из них свободный, поскольку они подчиняются тождеству

$$p^2+q^2=1 \tag{4.44}$$

Недостающее одно уравнение для определения этих параметров получается из условия обращения в нуль некоторого элемента новой матрицы $A^{(1)}$.В зависимости от выбора этого элемента строятся различные алгоритмы метода вращений.

Одним из таких алгоритмов является последовательное обращение в нуль всех ненулевых элементов, лежащих вне трех диагоналей исходной симметрической матрицы. Это так называемый *прямой метод вращений.* В соответствии с этим методом обращение в нуль элементов матрицы производится последовательно, начиная с элементов первой строки (и первого столбца, так как матрица симметрическая).

Процесс вычислений поясним с использованием схематического изображения матрицы (Рисунок 4.6). Точками отмечены элементы матрицы. Наклонные линии указывают три диагонали матрицы, элементы на которых после окончания расчета отличны от нуля. Алгоритм решения задачи нужно построить таким образом, чтобы все элементы по одну сторону от этих трех диагоналей обратились в нуль; тогда симметрично расположенные элементы также станут нулевыми. Обращение элементов в нуль можно выполнять, например, в следующей последовательности: a_{13}, a_{14}, …, a_{1n}, a_{24}, a_{25}, …, a_{2n}, …, a_{n-2n}.

Рассмотрим сначала первый шаг данного метода, состоящий в обращении в нуль элемента a_{13} (и автоматически a_{31}). Для осуществления элементарного вращения нужно выбрать две оси — i-ю и j-ю, так чтобы элемент a_{13} оказался в строке или столбце с номером i или *j*. Положим *i* = 2, *j*=3 и умножим матрицу $A^{(0)}$ справа на матрицу вращения P_{23} и слева на транспонированную матрицу P_{23}^T. Получим новые значения элементов матрицы, которые вычисляются по формулам (5.42), (5.43). Полагая в них l=1 и *m*=3, находим

$$a_{13}^{(1)} = b_{13} = -a_{12} q + a_{13} p = 0.$$

Учитывая тождество p2+q2=1 (4.44), получаем систему уравнений для определения параметров *p, q:*

$$\begin{aligned} a_{13} p - a_{12} q &= 0, \\ p^2 + q^2 &= 1. \end{aligned}$$

Решая эту систему, находим

$$p = \frac{a_{12}}{\sqrt{a_{12}^2 + a_{13}^2}},\ q = \frac{a_{12}}{\sqrt{a_{12}^2 + a_{13}^2}}.$$

Используя эти параметры p, q, можно по формулам (5.42), (5.43) вычислить значения элементов, стоящих в строках и столбцах с номерами i=2,3; j=2,3 (остальные элементы исходной матрицы не изменились).

Аналогично, выбирая для элементарного вращения i-ю и j-ю оси, можно добиться нулевого значения любого элемента $a_{i-1,j}^{(k)}$ на k-м шаге. В этом случае строится матрица вращения $P_{i,j,}$параметры которой вычисляются по формулам, полученным из условия равенства нулю элемента $a_{i-1,j}^{(k)}$ и (5.44). Эти формулы имеют вид

$$p = \frac{a_{i-1,i}^{(k-1)}}{\sqrt{\left(a_{i-1,i}^{(k-1)}\right)^2 + \left(a_{i-1,j}^{(k-1)}\right)^2}},\ q = \frac{a_{i-1,j}^{(k-1)}}{\sqrt{\left(a_{i-1,i}^{(k-1)}\right)^2 + \left(a_{i-1,j}^{(k-1)}\right)^2}}.$$

Учитывая найденные значения параметров *p, q,* можно по формулам (5.42), (5.43) найти элементы преобразованной матрицы. Для иллюстрации вновь обратимся к Рисунок 4.6. Вертикальными линиями показаны столбцы с номерами *i* и *j*, соответствующими осям элементарного вращения, горизонтальными — строки с теми же номерами.

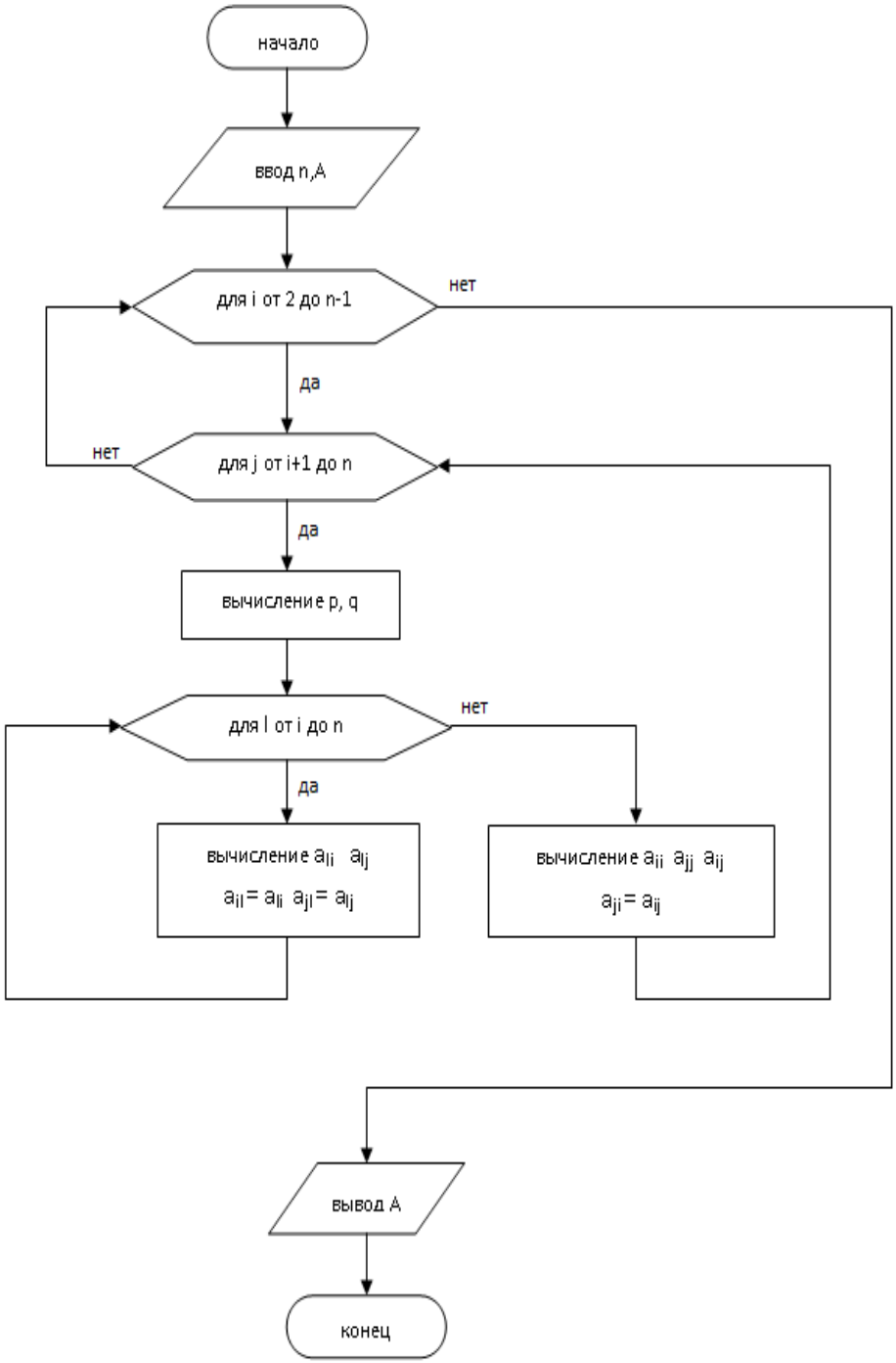

Рисунок 4.6

На рассматриваемом шаге матрица преобразуется таким образом, чтобы отмеченные крестиками элементы обратились в нуль. Элементарное вращение (5.41) на каждом шаге требует пересчета всех элементов отмеченных столбцов и строк. Учитывая симметрию, можно вычислить лишь все элементы столбцов, а элементы получаются из условий симметрии. Исключение составляют лишь элементы, расположенные на пересечениях этих строк и столбцов. Они изменяются на каждом из двух этапов выполняемого шага. Таким образом, на каждом шаге преобразования симметрической матрицы для вычисления элементов столбцов используются формулы (5.42), а элементы, находящиеся на пересечениях изменяемых строк и столбцов, пересчитываются еще по формулам (5.43). При этом полученные ранее нулевые элементы не изменяются. Алгоритм приведения симметрической матрицы к трехдиагональному виду с помощью прямого метода вращений представлен на рисунке.

Собственные значения полученной трехдиагональной матрицы будут также собственными значениями исходной матрицы. Собственные векторы x_i исходной матрицы не равны непосредственно собственным векторам y_i трехдиагональной матрицы, а вычисляются с помощью соотношений

$$x_i = P_{23}P_{24}\ldots P_{n-1,n}y_i \qquad (4.45)$$

3.3 Трехдиагональные матрицы.

Симметрическую матрицу можно привести с помощью преобразований подобия к трехдиагональному виду. Кроме того, трехдиагональные матрицы представляют самостоятельный интерес, поскольку они встречаются в вычислительной практике, и нередко требуется находить их собственные значения и собственные векторы. Рассмотрим трехдиагональную матрицу вида

$$A = \begin{pmatrix} b_1 & c_1 & & & & 0 \\ a_1 & b_2 & c_2 & & & \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ & & & a_{n-1} & b_{n-1} & c_{n-1} \\ 0 & & & & a_n & b_n \end{pmatrix}. \qquad (4.46)$$

Здесь элементы $b_1,\ b_2, \ldots, b_n$ расположены вдоль главной диагонали; $c_1,\ c_2, \ldots, c_{n-1}$ - над ней; $a_2,\ a_3, \ldots, a_n$ - под ней.

Для нахождения собственных значений нужно приравнять нулю определитель $D_n(\lambda)$ = dot(A - λ E), или

$$D_n(\lambda) = \begin{vmatrix} b_1-\lambda & c_1 & & & & 0 \\ a_2 & b_2-\lambda & c_2 & & & \\ & & & & & \\ & & & a_{n-1} & b_{n-1} & c_{n-1} \\ 0 & & & & a_n & b_n \end{vmatrix} = 0. \qquad (4.47)$$

Произвольный определитель *n*-го порядка можно выразить через *n* миноров *(n*-1)-го порядка путем разложения его по элементам любой строки или любого столбца. Разложим определитель (5.47) по элементам последней строки, в которой всего два ненулевых элемента. Получим

$$D_n(\lambda) = (b_n-\lambda)D_{n-1}(\lambda) - a_n M_{n-1}(\lambda),$$

$$M_{n-1}(\lambda) = \begin{vmatrix} b_1-\lambda & c_1 & & & 0 \\ a_2 & b_2-\lambda & c_2 & & \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & & & a_{n-1} & c_{n-1} \end{vmatrix}. \qquad (4.48)$$

Поскольку минор $M_{n-1}(\lambda)$ содержит в последнем столбце лишь один ненулевой элемент c_{n-1} то, разлагая его по элементам этого столбца, получаем

$$M_{n-1}(\lambda) = c_{n-1}D_{n-2}(\lambda).$$

Подставляя это выражение в формулу (5.48), получаем рекуррентные соотношения, выражающие минор высшего порядка через миноры двух низших порядков:

$$D_n(\lambda) = (b_n-\lambda)D_{n-1}(\lambda) - a_n c_{n-1}D_{n-2}(\lambda). \qquad (4.49)$$

Положим $D_0(\lambda)$= 1. Минор первого порядка равен элементу a_{11} определителя, т. е. в данном случае $D_1(\lambda)$=$b_1 - \lambda$. Проверим, с учетом значений $D_0(\lambda)$, $D_1(\lambda)$ правильность формулы (4.49) при *n* = 2;

$$D_2(\lambda) = (b_2-\lambda)D_1(\lambda) - a_2c_1D_o(\lambda) = (b_2-\lambda)(b_1-\lambda) - a_2c_1. \qquad (4.50)$$

Вычисляя минор второго порядка определителя (5.47), убеждаемся в справедливости выражения (5.50). Таким образом, используя рекуррентные соотношения (5.49), можно найти выражение для характеристического многочлена $D_n(\lambda)$ для любого $n \geq 2$. Вычисляя корни этого многочлена, получаем собственные значения $\lambda_1,\ \lambda_2, \ldots, \lambda_n$ трехдиагональной матрицы (5.46).

Будем считать, что собственные значения $\lambda_1,\ \lambda_2, \ldots, \lambda_n$ матрицы (5.46) вычислены. Найдем соответствующие им собственные векторы. Для любого собственного значения собственный вектор находится из системы уравнений (5.36).

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

Перейдем от матричной формы записи этой системы к развернутой *(A*- матрица вида (5.46), x-{ x_1, x_2, ..., x_n}):

$$
\begin{aligned}
&(b_1-\lambda)x_1+c_1x_2 &&=0,\\
&a_2x_1+(b_2-\lambda)x_2+c_2x_3 &&=0,\\
&\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots &&\\
&a_{n-1}x_{n-2}+(b_{n-1}-\lambda)x_{n-1}+c_{n-1}x_n &&=0,\\
&a_nx_{n-1}+(b_n-\lambda)x_n &&=0.
\end{aligned}
\qquad (4.51)
$$

Матрица системы (5.51) вырожденная, поскольку ее определитель (5.47) равен нулю. Можно показать, что если $c_i \neq 0$ (i = 1,2,..., n-1), то последнее уравнение системы (5.51) является следствием остальных уравнений. Действительно, если отбросить первый столбец и последнюю строку в матрице A, то вместо (5.47) получится определитель вида

$$
\begin{vmatrix}
c_1 & & & 0\\
b_2-\lambda & c_2 & & \\
 & & & \\
0 & a_{n-1} & b_{n-1}-\lambda & c_{n-1}
\end{vmatrix} = c_1c_2\dots c_{n-1} \neq 0. \qquad (4.52)
$$

Следовательно, все строки с первой по (n-1)-ю линейно независимы, последнее уравнение системы (5.51) — следствие остальных, и одно неизвестное этой системы является свободным. Отбрасывая последнее уравнение системы (5.51), записываем ее в виде

$$
\begin{aligned}
&c_1x_2 &&=-(b_1-\lambda)x_1,\\
&(b_2-\lambda)x_2+c_2x_3 &&=-a_2x_1,\\
&\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots &&\\
&a_{n-1}x_{n-2}+(b_{n-1}-\lambda)x_{n-1}+c_{n-1}x_n &&=0.
\end{aligned}
\qquad (4.53)
$$

Считая компоненту x_1 свободным неизвестным и полагая ее равной любому ненулевому значению (иначе при $x_1 = 0$ остальные компоненты также будут нулевыми), можно из системы (5.53) найти последовательно все остальные компоненты: из первого уравнения легко вычислить x_2, из второго x_3 и т, д., из последнего x_n. Поскольку определитель (5.52) этой системы отличен от нуля, то она имеет единственное решение. Описанным способом могут быть найдены собственные векторы, соответствующие всем собственным значениям $\lambda_1, \lambda_2, \dots, \lambda_n$ трехдиагональной матрицы (5.46).

Если трехдиагодальная матрица получена в результате последовательности преобразований подобия исходной симметричной матрицы, то, как уже отмечалось, все собственные значения трехдиагональной матрицы являются одновременно собственными значениями исходной матрицы, а собственные векторы пересчитываются по формулам (5.45). При этом вычислять произведения матриц $P_{n-1,n,} \dots, P_{24}, P_{23}$ на собственные векторы трехдиагональной матрицы нецелесообразно, поскольку при умножении матрицы P_{ij}, на вектор x изменяются только две его компоненты: x_i и x_j. Поэтому в качестве этих компонент берем значения px_i -qx_j и qx_i + px_j, что сокращает объем вычислений по сравнению с умножением матрицы P_{ij}, на вектор x.

Глава 5
ПРИБЛИЖЕНИЕ ФУНКЦИЙ

§1.Точечная аппроксимация. Равномерное приближение.

1.1 Точечная аппроксимация.

Одним из основных типов точечной аппроксимации является интерполирование. Оно состоит в следующем: для данной функции

$$y=f(x) \qquad (5.1)$$

строим многочлен (6.1), принимающий в заданных точках x те же значения y_i, что и функция f(x), т.е i=0,1,....,n.

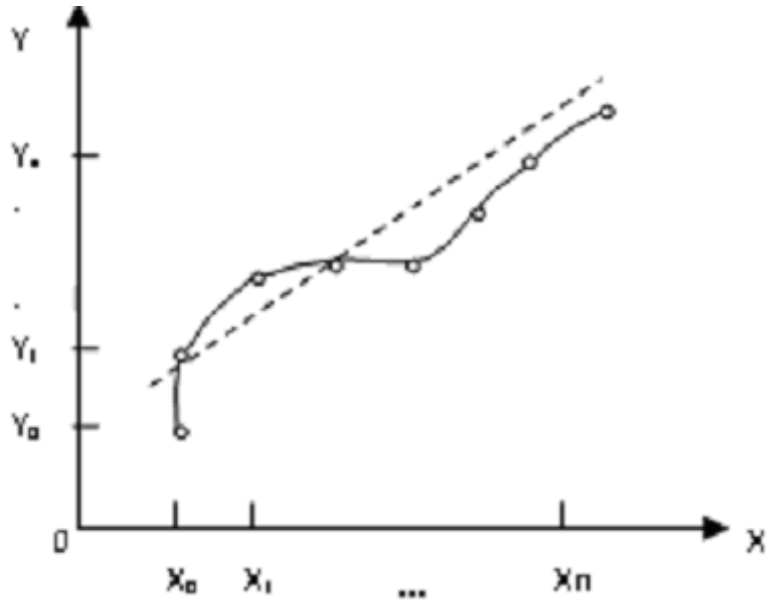

Рисунок 5.1

При этом предполагается, что среди значений x_i нет одинаковых, т.е. $x_i \neq x_k$ при $i \neq k$.

Точки x_i называются узлами интерполяции, а многочлен $\varphi(x)$ - интерполяционным многочленом.

Таким образом, близость интерполяционного многочлена к заданной функции состоит в том, что их значения совпадают на заданной системе точек (рисунок 6.1, сплошная линия).

Максимальная степень интерполяционного многочлена m=n, в этом случае говорят о глобальной интерполяции, поскольку один многочлен

$$\varphi(x) = a_0 + a_1 x + ... + a_n x^n \qquad (5.2)$$

используется для интерполяции функции f(x) на всем рассматриваемом интервале изменения аргумента x. Коэффициенты a_i многочлена находятся из системы (6.2). Можно показать, что при $x_i \neq x_k$ $(i \neq k)$ эта система имеет единственное решение.

Интерполяционные многочлены могут строиться отдельно для разных частей рассматриваемого интервала изменения x. В этом случае имеем кусочную (или локальную) интерполяцию.

Как правило интерполяционные многочлены используется для аппроксимации функции в промежуточных точках между крайними узлами интерполяции, т.е. при $x_0 < x < x_n$. Однако они используются и для приближенного вычисления функции вне рассматриваемого отрезка ($x < x_0$, $x > x_n$). Это приближение называют экстраполяцией.

При интерполировании основным условием является прохождение графика интерполяционного многочлена через данные значения функции в узлах интерполяции. Однако в ряде случаев выполнение этого условия затруднительно или даже нецелесообразно.

Например, при большом числе узлов интерполяции получается высокая степень многочлена (6.2) в случае глобальной интерполяции, т.е. когда нужно иметь один интерполяционный многочлен для всего интервала изменения аргумента. Кроме того, табличные данные могли быть получены путем измерений и содержать ошибки. Построение аппроксимирующего многочлена с условием обязательного прохождения его графика через эти экспериментальные точки означало бы тщательное повторение допущенных при измерениях ошибок. Выход из этого положения может быть найден выбором такого многочлена, график которого

проходит близко от данных точек (рисунок 6.1, штриховая линия). Понятие «близко» уточняется при рассмотрении разных видов приближения.
Одним из таких видов является среднеквадратичное приближение функции с помощью многочлена (6.1). При этом $m \le n$; случай m=n соответствует интерполяции. На практике стараются подобрать аппроксимирующий многочлен как можно меньшей степени (как правило, m=1, 2. 3).
Мерой отклонения многочлена $\varphi(x)$ от заданной функции f(x) на множестве точек (x_i,y_i) (i=0,1, 2, …, n) при среднеквадратичном приближении является величина S, равная сумме квадратов разностей между значениями многочлена и функции в данных точках:

$$S = \sum_{i=0}^{n} [\varphi(x_i) - y_i]^2 \qquad (5.3)$$

Для построения аппроксимирующего многочлена нужно подобрать коэффициенты a_0, a_1, …, a_m так, чтобы величина S была наименьшей. В этом состоит метод наименьших квадратов.

1.2 Равномерное приближение.

Во многих случаях, особенно при обработке экспериментальных данных, среднеквадратичное приближение вполне приемлемо, поскольку оно сглаживает некоторые неточности функции f(x) и дает достаточно правильное представление о ней. Иногда, однако, при построении приближения ставится более жесткое условие: требуется, чтобы во всех точках некоторого отрезка [a. b] отклонение многочлена $\varphi(x)$ от функции f(x) было по абсолютной величине меньшим заданной величины $\varepsilon > 0$:

$$\left| f(x) - \varphi(x) \right| < \varepsilon, a \le x \le b.$$

В этом случае говорят, что многочлен $\varphi(x)$ равномерно аппроксимирует функцию f(x) с точностью ε на отрезке [a, b].
Введем понятие абсолютного отклонения Δ многочлена $\varphi(x)$ от функции f(x) на отрезке [a, b]. Оно равно максимальному значению абсолютной величины разности между ними на данном отрезке:

$$\Delta = \max_{a \le x \le b} \left| f(x) - \varphi(x) \right|.$$

По аналогии можно ввести понятие среднеквадратичного отклонения $\overline{\Delta} = \sqrt{S/n}$ при среднеквадратичном приближении функций.
Возможность построения многочлена, равномерно приближающего данную функцию, следует из теоремы Вейерштрасса об аппроксимации:
Теорема. Если функция f(x) непрерывна на отрезке [a, b], то для любого $\varepsilon > 0$ существует многочлен $\varphi(x)$ степени $m = m(\varepsilon)$, абсолютное отклонение которого от функции f(x) на отрезке [a, b] меньше ε.

§2. Многочлены Чебышева. Вычисление многочленов. Рациональные приближения.

2.1 Многочлены Чебышева.

Одним из способов совершенствования алгоритма вычислений, позволяющих более равномерно распределить погрешность по всему интервалу, является использование многочленов Чебышева.
Многочлен Чебышева

$$T_n(x) = \frac{1}{2}[(x + \sqrt{x^2 - 1})^n + (x - \sqrt{x^2 - 1})^n], \; -1 \le x \le 1 \qquad (5.4)$$

n=0,1,…
Легко показать, что (6.4) действительно является многочленом: при возведении в степень и последующих преобразованиях члены, содержащие корни, уничтожаются. Приведем многочлены Чебышева, полученные по формуле (6.3) при n=0,1,2,3

$$T_0(x) = 1, \; T_1(x) = x, \; T_2(x) = 2x^2 - 1, \; T_3(x) = 4x^3 - 3x.$$

Для вычисления многочлена Чебышева можно воспользоваться рекуррентным соотношением

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \qquad (5.5)$$

n=1,2,…
В ряде случаев важно знать коэффициент a_n при старшем члене многочлена Чебышева степени n

$$T_n(x) = a_0 + a_1 x + \ldots + a_n x^n.$$

Разделив этот многочлен на x^n, найдем

$$a_n = \frac{T_n(x)}{x^n} - \frac{a_0}{x^n} - \ldots - \frac{a_{n-1}}{x}.$$

Перейдем к пределу при $x \to \infty$ и воспользуемся формулой (6.4), получим

$$a_n = \lim_{x\to\infty}\frac{T_n(x)}{x^n} = \frac{1}{2}\lim_{x\to\infty}[(1+\sqrt{1-\frac{1}{x^2}})^n + ?(1-\sqrt{1-\frac{1}{x^2}})^n] = 2^{n-1}.$$

Многочлены Чебышева можно так же представить в тригонометрической форме:

$$T_n(x) = \cos(n\arccos x) \qquad (5.6)$$

n =0,1,…

С помощью этих выражений могут быть получены формулы (6.5).

Нули (корни) многочленов Чебышева на отрезке [-1,1] определяются формулой

$$x_k = \cos\frac{2k-1}{2n}\pi \text{ , k=1,2,…,n.}$$

Они расположены неравномерно на отрезке и сгущаются к его концам.

Вычисляя экстремумы многочлена Чебышева по обычным правилам (с помощью производных), можно найти его максимумы и минимумы:

$$x_k = \cos(k\pi/n) \text{ , k=1,2,…,n-1.}$$

В этих точках многочлен принимает поочередно значения $T_n(x_k)$=-(+)1, т.е. все максимумы равны 1, а минимумы -1. На границах отрезка значения многочленов Чебышева равны –(+)1.

Многочлены Чебышева широко используются при аппроксимации функций. Рассмотрим их применение для улучшения приближения функций с помощью степенных рядов, а именно для более равномерного распределения погрешностей аппроксимации по заданному отрезку [-π/2, π/2].

Отрезок [-π/2, π/2] является не совсем удобным при использовании многочленов Чебышева, поскольку они обычно рассматриваются на стандартном отрезке [-1,1]. Первый отрезок легко привести ко второму заменой переменной x на π/2. В этом случае ряд для аппроксимации синуса на отрезке [-1,1]примет вид:

$$\sin\frac{\pi x}{2} = \frac{\pi x}{2} - \frac{1}{3!}(\frac{\pi x}{2})^2 + \frac{1}{5!}(\frac{\pi x}{2})^5 - ... \qquad (5.7)$$

При использовании этого ряда погрешность вычисления функции в окрестности концов отрезка $x=\pm 1$ существенно возрастает и становится значительно больше, чем в окрестности точки x=0. Если вместо (6.6) использовать ряд

$$\sin(\frac{\pi x}{2}) = c_0 + c_1T_1(x) + c_2T_2(x) + ...,$$

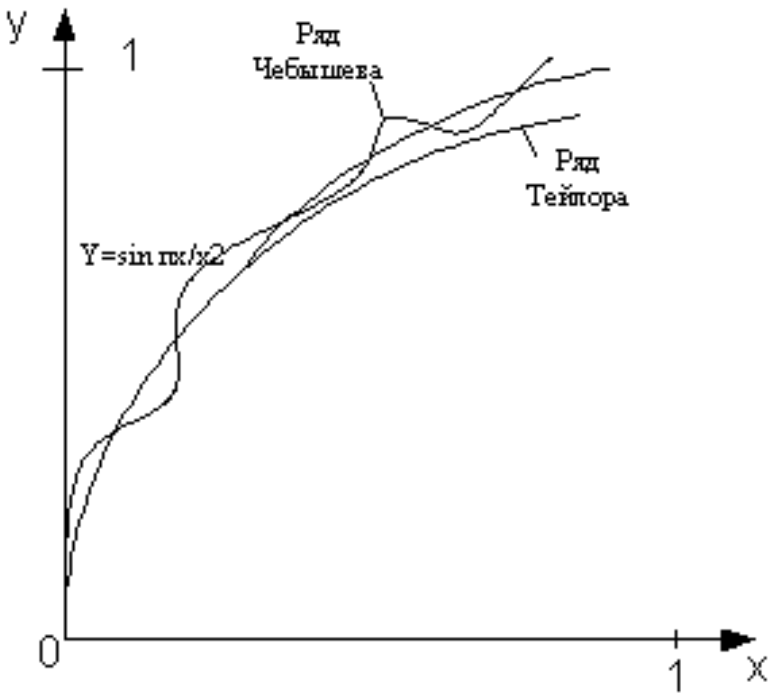

Рисунок 5.2

члены которого являются многочлены Чебышева, то погрешность будет распределена по всему отрезку (рисунок 6.2). В частности при использовании многочленов Чебышева до девятой степени включительно погрешность находится в интервале $(-5\div 5)*10^{-9}$. Причем погрешность ряда Тейлора для этой задачи на концах отрезка составляет $4*10^{-6}$.

Нахождение коэффициентов ряда Чебышева довольно сложно и здесь рассматриваться не будет. На практике часто используют многочлены Чебышева для повышения точности аппроксимации функций с помощью ряда Тейлора.

Пусть частичная сумма ряда Тейлора, представленная в виде многочлена, используется для приближения функции f(x) на стандартном отрезке [—1, 1], т. е.

$$f(x) \approx a_0 + a_1 x + \ldots + a_n x^n \tag{5.8}$$

Если рассматриваемый отрезок [a, b] отличается от стандартного, то его всегда можно привести к стандартному заменой переменной

$$t = \frac{a+b}{2} + \frac{b-a}{2} x_2 \qquad -1 \le x \le 1.$$

Многочлен Чебышева $T_n(x)$ можно записать в виде

$$T_n(x) = b_0 + b_1 x + b_2 x^2 + \ldots + 2^{n-1} x^n.$$

Отсюда получаем

$$x^n = -2^{1-n}(b_0 + b_1 x + \ldots + b_{n-1} x^{n-1}) + 2^{1-n} T_n(x) \tag{5.9}$$

Если отбросить последний член, то допущенную при этом погрешность Δ легко оценить: $|\Delta| \le 2^{1-n}$, поскольку $|T_n(x)| \le 1$. Таким образом, из (6.8) получаем, что x^n есть линейная комбинация более низких степеней x. Подставляя эту линейную комбинацию в (6.7), приходим к многочлену степени n- 1 вместо многочлена степени n. Этот процесс может быть продолжен до тех пор, пока погрешность не превышает допустимого значения.

Используем эту процедуру для повышения точности аппроксимации функции с помощью ряда (6.5). Будем учитывать члены ряда до 11-й степени включительно. Вычисляя коэффициенты при степенях x, получаем

$$\sin(\frac{\pi x}{2}) \approx 1.5707963x - 0.64596410x^3 + 0.79692626x^5 - 0.0046817541x^7 + 0.00016044118x^9 - 0.0000035988432x^{11} \tag{5.10}$$

Многочлен Чебышева 11-й степени имеет вид

$$T_{11} = 1024x^{11} - 2816x^9 + 2816x^7 - 1232x^5 + 220x^3 - 11x.$$

Выразим отсюда x^{11} через более низкие степени:

$$x^{11} = 2\text{-}10\,(11x - 220x^3 + 1232x^5 - 2816x^7 + 2816x^9 + T_{11}).$$

Подставляя в (6.10) вместо x^{11} правую часть этого равенства и вычисляя новые значения коэффициентов, получаем:

$$\sin(\frac{\pi x}{2}) \approx 1.5707962x - 0.64596332x^3 + 0.079688296x^5 - 0.0046718573x^7 + 0.00015054436x^9 - 0.00000000351T_{11} \tag{5.11}$$

Отбрасывая последний плен этого разложения, мы допускаем погрешность $|\Delta| \le 3.51 \cdot 10^{-9}$. Из-за приближенного вычисления коэффициентов при степенях x реальная погрешность больше. Здесь она оценивается величиной $|\Delta| \le 8 \cdot 10^{-9}$.Эта погрешность немного больше, чем для многочлена Чебышева ($5*10^{-9}$), и значительно меньше, чем для ряда Тейлора ($4*10^{-6}$).

Процесс модификации приближения можно продолжить. Если допустимое значение погрешности больше, чем при использовании выражения (6.12) (без последнего члена с T_{11}), то x^9 можно заменить многочленом седьмой степени, а член с T_9 отбросить; так продолжать до тех пор, пока погрешность остается меньше допустимой.

В заключение приведем некоторые формулы, необходимые при использовании многочленов Чебышева.

1. Многочлены Чебышева:

$$T_n(\mathrm{x}) = \frac{1}{2}\left[(x + \sqrt{x^2 - 1})^n + (x - \sqrt{x^2 - 1})^n\right] = \cos(n \arccos x),$$

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x),\ \mathrm{n}=1,2,\ldots,$$

$$T_0(x) = 1,$$

$$T_1(x) = x,$$

$$T_2(x) = 2x^2 - 1,$$

$$T_3(x) = 4x^3 - 3x,$$

$$T_4(x) = 8x^4 - 8x^2 + 1,$$
$$T_5(x) = 16x^5 - 20x^3 + 5x,$$
$$T_6(x) = 32x^6 - 48x^4 + 18x^2 - 1,$$
$$T_7(x) = 64x^7 - 112x^5 + 56x^3 - 7x,$$
$$T_8(x) = 128x^8 - 256x^6 + 160x^4 - 32x^2 + 1,$$
$$T_{10}(x) = 512x^{10} - 1280x^8 + 1120x^5 - 400x^4 + 50x^2 - 1,$$
$$T_{11}(x) = 1024x^{11} - 2816x^9 + 2816x^6 - 1232x^5 + 220x^3 - 11x,$$

2. Представление степеней x через многочлены $T_n(x)$:

$$x^0 = 1 = T_{01},$$
$$x = T_{11},$$
$$x^2 = \frac{1}{2}(T_0 + T_2),$$
$$x^3 = \frac{1}{4}(3T_1 + T_3),$$
$$x^4 = \frac{1}{8}(3T_0 + 4T_2 + T_4)$$
$$x^5 = \frac{1}{16}(10T_1 + 5T_3 + T_5)$$
$$x^6 = \frac{1}{32}(10T_1 + 5T_3 + T_5)$$
$$x^7 = \frac{1}{64}(35T_1 + 21T_3 + 7T_5 + T_7)$$
$$x^8 = \frac{1}{128}(35T_0 + 56T_2 + 28T_4 + 8T_6 + T_8)$$
$$x^9 = \frac{1}{256}(126T_1 + 84T_3 + 36T_5 + 9T_7 + T_9)$$
$$x^{10} = \frac{1}{512}(126T_0 + 210T_2 + 120T_4 + 45T_6 + 10T_8 + T_{10})$$
$$x^{11} = \frac{1}{1024}(462T_1 + 330T_3 + 165T_5 + 55T_7 + 11T_9 + T_{11})$$

3. Выражение x^n через более низкие степени

$x=T_{11}$

$$x^2 = \frac{1}{2}(1 + T_2),$$
$$x^3 = \frac{1}{4}(3x + T_3),$$
$$x^4 = \frac{1}{8}(8x^2 - 1 + T_4),$$
$$x^5 = \frac{1}{16}(20x^3 - 5x + T_5),$$
$$x^6 = \frac{1}{32}(48x^4 - 18x^2 + 1 + T_6),$$
$$x^7 = \frac{1}{64}(112x^5 - 56x^3 + 7x + T_7),$$
$$x^8 = \frac{1}{128}(256x^6 - 160x^4 + 32x^2 - 1 + T_8),$$

$$x^9 = \frac{1}{256}(576x^7 - 432x^5 + 120x^3 - 9x + T_9),$$

$$x^{10} = \frac{1}{512}(1280x^8 - 1120x^6 + 400x^4 - 50x^2 - 1 + T_{10}),$$

$$x^{11} = \frac{1}{1024}(2816x^9 - 2816x^7 + 1232x^5 - 220x^3 - 11x + T_{11})$$

2.2 Вычисление многочленов.

При аппроксимации функций, а также в некоторых других задачах приходится вычислять значения многочленов вида

$$P(x) = a_0 + a_1x + a_2x^2 + ... + a_nx^n \tag{5.12}$$

Если проводить вычислении «в лоб», т. е. находить значения каждого члена и суммировать их, то при больших n потребуется выполнить большое число операций($n^2 + n/2$ умножений и n сложений). Кроме того, это может привести к потере точности за счет погрешностей округления. В некоторых частных случаях, как это сделано при вычислении синуса, удается выразить каждый последующий член через предыдущий и таким образом значительно сократить объем вычислений.

Анализ многочлена (6.13) в общем случае приводит к тому, что для исключения возведения x в степень в каждом члене многочлен целесообразно переписать в виде

$$P(x) = a_0 + x(a_1 + x(a_2 + ... + x(a_{n-1} + xa_n) + ...)). \tag{5.13}$$

Прием, с помощью которого многочлен представляется в таком виде, называется схемой Горнера. Этот метод требует n умножений на n сложении. Использование схемы Горнера для вычисления значений многочленов не только экономит машинное время, но и повышает точность вычислений за счет уменьшения погрешностей округления.

2.3 Рациональные приближения.

Рассмотрим другой вид аппроксимации функций — с помощью дробно- рационального выражения. Функцию представим в виде отношения двух многочленов некоторой степени. Пусть, например, это будут многочлены третьей степени, т. е. представим функцию f(x) d виде дробно- рационального выражения:

$$f(x) = \frac{b_0 + b_1x + b_2x^2 + b_3x^3}{1 + c_1x + c_2x^2 + c_3x^3} \tag{5.14}$$

Значение свободного члена в знаменателе $c_0 = 1$ не нарушает общности этого выражения, поскольку при $c_0 \neq 1$ числитель и знаменатель можно разделить на c_0. Перепишем выражение (6.13) в виде

$$b_0 + b_1x + b_2x^2 + b_3x^3 = (1 + c_1x + c_2x^2 + c_3x^3) f(x).$$

Используя разложение функции f(x) в ряд Тейлора:

$$f(x) = a_0 + a_1x + a_2x^2 + ... \tag{5.15}$$

и учитывая члены до шестой степени включительно, получаем

$$b_0 + b_1x + b_2x^2 + b_3x^3 = (1 + c_1x + c_2x^2 + c_3x^3) \times (a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5 + a_6x^6)$$

Преобразуем правую часть этого равенства, записав ее разложение по степеням x:

$$b_0 + b_1x + b_2x^2 + b_3x^3 = a_0 + x(a_1 + a_0c_1) + x^2(a_2 + a_1c_1 + a_0c_2) + x^3(a_3 + a_2c_1 + a_1c_2 + a_0c_3) +$$

$$+ x^4(a_4 + a_3c_1 + a_2c_2 + a_1c_3) + x^5(a_5 + a_4c_1 + a_3c_2 + a_2c_3) + x^6(a_6 + a_5c_1 + a_4c_2 + a_3c_3).$$

Приравнивая коэффициенты при одинаковых степенях x левой и правой частях, получаем следующую систему уравнений:

$$b_0 = a_0,$$

$$b_1 = a_1 + a_0c_1,$$

$$b_2 = a_2 + a_1c_1 + a_0c_2,$$

$$b_3 = a_3 + a_2c_1 + a_1c_2 + a_0c_3,$$

$$0 = a_4 + a_3c_1 + a_2c_2 + a_1c_3,$$

$$0 = a_5 + a_4c_1 + a_3c_2 + a_2c_3,$$

$$0 = a_6 + a_5c_1 + a_4c_2 + a_3c_3.$$

Решив эту систему, найдем коэффициенты b_0, b_1, b_2, b_3, c_1, c_2, c_3, необходимо для аппроксимации (6.13).

Пример.

Рассмотрим рациональное приближение для функции $f(x)=\sin(\frac{\pi x}{2})$. Воспользуемся представлением (6.13), которое в данном случае упрощается, поскольку функция sin x нечетная. В частности в частности, в числителе можем оставить только члены с нечетными степенями л, а в знаменателе — с четными; коэффициенты при других степенях x равны нулю: $b_1 = b_2 = c_1 == c_3$=0,

Коэффициенты b_1, b_2, c_2, найдем из системы уравнений (6.13), причем значения коэффициентов a_0, a_1, …, a_6 разложения функции в ряд Тейлора (6.12) $P(x)=a_0+a_1x+a_2x^2+...+a_nx^n$ (5.12)можем взять из выражения (6.4), т. е.

$$a_0=0,\qquad a_1=\frac{\pi}{2},\qquad a_2=0,\qquad a_3=-\frac{\pi^3}{8\cdot 3}!,\quad a_4=0,\qquad a_5=\frac{\pi^5}{32\cdot 5!},\ a_6=0.$$

Система уравнений в данном случае примет вид

$$b_1=\frac{\pi}{2},\qquad 8\qquad 0=\frac{\pi^5}{32\cdot 5!}-\frac{\pi^3}{8\cdot 3!}c_2.$$

Отсюда находим $b_1=\frac{\pi}{2},\ b_3=-\frac{7\pi^3}{480},\ c_2=\frac{\pi^2}{80}$.

Таким образом, дробно- рациональное приближение (6.11) для функции $f(x)=\sin(\frac{\pi x}{2})$ примет вид

$$\sin(\frac{\pi x}{2})=\frac{(\pi/2)x-(7\pi^3/480)x^3}{1+(\pi^2/80)x^2} \tag{5.16}$$

Это приближение по точности равносильно аппроксимации (6.4) с учетом членов до пятого порядка включительно.

На практике с целью экономии числа операций выражение (6.11) представляется в виде цепной дроби. Представим в таком виде дробно-рациональное выражение (6.14). Сначала перепишем это выражение, вынося за скобки коэффициенты при x^3 и x^2. Получим

$$\sin(\frac{\pi x}{2})=-\frac{7\pi}{6}\frac{x^3-(60/7)(2/\pi)^2x}{x^2+20(2/\pi)^2}.$$

Разделим числитель на знаменатель по правилу деления многочленов и введем обозначения для коэффициентов.

Получим

$$\sin(\frac{\pi x}{2})=k_1(x+\frac{k_2x}{x^2+k_3}),$$

$$k_1=-\frac{7\pi}{6},\qquad k_2=-\frac{200}{7}(\frac{2}{\pi})^2,\ k_3=20(\frac{2}{\pi})^2.$$

Полученное выражение можно записать в виде

$$\sin(\frac{\pi x}{2})=k_1\left(x+\frac{k_2}{x+\frac{k_3}{x}}\right) \tag{5.17}$$

Для вычисления значения функции по этой формуле требуется намного меньше операций (два деления, два сложения, одно умножение), чем для вычисления с помощью выражения (6.15) или усеченного ряда Тейлора (6.4) (далее с использованием правила Горнера).

Приведем формулы для приближения некоторых элементарных функций с помощью цепных дробей, указывая интервалы изменения аргумента и погрешности Δ:

$$e^x=1+\cfrac{x}{\cfrac{x}{2}+\cfrac{k_0+k_1x^2}{1+k_2x^2}} \tag{5.18}$$

$$k_0 = 1.0000000020967\,,\ -\frac{1}{2}\ln 2 \le x \le \frac{1}{2}\ln 2\,,\ |\Delta| \le 10^{-10};$$

$$\ln(1+x) = k_0 + \cfrac{x}{k_1 + \cfrac{x}{k_2 + \cfrac{x}{k_3 + \cfrac{x}{k_4 + \cfrac{x}{k_5}}}}} \tag{5.19}$$

$$k_0 = 0.0000000894\,,\ k_1 = 1.0000091365\,,$$
$$k_2 = 2.0005859000\,,\ k_3 = 3.0311932666\,,$$
$$k_4 = 1.0787748225\,,\ k_5 = 8.8952784060\,,$$
$$0 \le x \le 1,\ |\Delta| \le 10^{-7};$$

$$tg\frac{\pi x}{4} = x\left(k_0 + \cfrac{x^2}{k_1 + \cfrac{x^2}{k_2 + \cfrac{x^2}{k_3}}}\right) \tag{5.20}$$

$$k_0 = 0.7853980289\,,\ k_1 = 6.1922344479\,,$$
$$k_2 = -0.6545887679\,,\ k_3 = 491.0013934779\,,$$
$$-1 \le x \le 1,\ |\Delta| \le 2\cdot 10^{-7}.$$

$$arctgx = x\left(k_0 + \cfrac{x^2}{k_1 + \cfrac{x^2}{k_2 + \cfrac{x^2}{k_3 + \cfrac{x^2}{k_4}}}}\right) \tag{5.21}$$

$$k_0 = 0.99999752\,,\ k_1 = -3.00064286\,,$$
$$k_2 = -0.55703890\,,\ k_3 = -17.03715998\,,$$
$$k_4 = -0.20556880\,,$$
$$-1 \le x \le 1,\ |\Delta| \le 2\cdot 10^{-7}.$$

§3. Интерполирование. Линейная и квадратичная интерполяция. Многочлен Лагранжа. Многочлен Ньютона. Кубические сплайны. Точность интерполяции.

3.1 Линейная квадратичная интерполяция.

Простейшим и часто используемым видом локальной интерполяции является линейная интерполяция. Она состоит в том, что заданные точки (x_i, y_i) (i=0,1,2,3,…,n)-соединяются прямолинейными отрезками, и функция f(x) приближается ломаной с вершинами в данных точках.

Уравнения каждого отрезка ломано разные. Поскольку имеется n интервалов (x_{i-1}, x_i), то для каждого из них в качестве уравнения интерполяционного многочлена не используется уравнение примой, проходящей через

две точки. В частности, для i-го интервала можно написать уравнение прямой, проходящей через точки (x_{i-1}, y_{i-1}) и (x_i, y_i), в виде

$$\frac{y-y_{i-1}}{y_i-y_{i-1}}=\frac{x-x_{i-1}}{x_i-x_{i-1}}.$$

Отсюда

$$y=a_ix+b_i,\ x_{i-1}\le x\le x_i \qquad (5.22)$$

, $$a_i=\frac{y_i-y_{i-1}}{x_i-x_{i-1}},\ b_i=y_{i-1}-a_ix_{i-1}.$$

Следовательно, при использовании линейной интерполяции сначала нужно определить интервал, в который попадает значение аргумента x, а затем подставить его в формулу (6.21) и найти приближенное значении е функции в этой точке.

Рассмотрим теперь случай квадратичной интерполяции. В качестве интерполяционной функции на отрезке [x_{i-1}, x_{i+1}] принимается квадратный трехчлен. Такую интерполяцию называют также параболической.

Уравнение квадратного трехчлена

$$y=a_ix^2+b_ix+c_i,\ x_{i-1}\le x\le x_{i+1} \qquad (5.23)$$

Содержит три неизвестных коэффициента a_i, b_i, c_i, для определения которых необходимо три уравнения. Ими служат условия прохождения параболы (6.22_через три точки (x_{i-1}, y_{i-1}), (x_i, y_i), (x_{i+1}, y_{i+1}). Эти условия можно записать в виде

$$\begin{aligned} a_ix_{i-1}^2+b_i\,x_{i-1}+c_i&=y_{i-1},\\ a_ix_i^2+b_ix_i+c_i&=y_i\\ a_ix_{i+1}^2+b_i\,x_{i+1}+c_i&=y_{i+1}\end{aligned} \qquad (5.24)$$

Алгоритм вычисления приближенного значения функции с помощью квадратичной интерполяций можно представить в виде блок-схемы, как и для случая линейной интерполяции. Вместо формулы (6.21) нужно использовать (6.22) с учетом решения системы линейных уравнений (6.23). Интерполяция для любой точки $x\in[x_0,x_n]$ проводится по трем ближайшим к ней узлам.

ПРИМЕР. Найти приближенное значение функции y = f(x) при x = 0.32, если известна следующая таблица ее значений:

x	0.15	0.30	0.40	0.55
y	2.17	3.63	5.07	7.78

Воспользуемся сначала формулой линейной интерполяции (6.22)

$$y=a_ix+b_i,\ x_{i-1}\le x\le x_i \qquad (5.22).$$

Значение X = 0.32 находится между узлами x_i = 0.30 и x_i = 0.40. В этом случае

$$a_i=\frac{y_i-y_{i-1}}{x_i-x_{i-1}}=\frac{5.07-3.63}{0.40-0.30}=14.4,$$

$$b_i=y_{i-1}-a_ix_{i-1}=3.63-14.4\cdot0.30=-0.69,$$

$$y\approx14.4x-0.69=14.4\cdot0.32-0.69=3.92.$$

Найдем теперь приближенное значение функции с помощью формулы квадратичной интерполяции (6.22). Составим систему уравнений (6.23) с учетом ближайших к точке x= 0.32 узлов: x_{i-1}=0.15, x_i =0.30, x_{i+1}=0.40. Соответственно y_{i-1} = 2.17, y_i = 3.63, y_{i+1}=5.07. Система (6.24) запишется в виде

$$0.15^2a_i+0.15b_i+c_i=2.17,$$

$$0.30^2a_i+0.30b_i+c_i=3.63,$$

$$0.40^2a_i+0.40b_i+c_i=5.07.$$

Решая эту систему, находим $a_i=18.67$, $b_i=1.33$, $c_i=1.55$. Искомое значение функции

$$y\approx18.67\cdot0.32^2+1.33\cdot0.32+1.55=3.89$$

3.2 Многочлен Лагранжа.

Перейдем к случаю глобальной интерполяции, т. е. построению интерполяционного многочлена, единого для всего отрезка $[x_0, x_n]$. При этом, естественно, график интерполяционного многочлена должен проходить через все заданные точки.

Запишем искомый многочлен в виде

$$\varphi(x) = a_0 + a_1 x + ... + a_n x^n \tag{5.25}$$

Из условий равенства значений этого многочлена в узлах x_i соответствующим заданным табличным значением y_i получим следующую систему уравнений для нахождения коэффициентов a0, a1, …,an:

$$\begin{aligned} a_0 + a_1 x_0 + ... + a_n x_0^n &= y_0 \\ a_0 + a_1 x_1 + ... + a_n x_1^n &= y_1, \\ &\cdots\cdots\cdots\cdots \\ a_0 + a_1 x_n + ... + a_n x_n^n &= y_n . \end{aligned} \tag{5.26}$$

Можно показать, что эта система имеет единственное решение, если среди узлов интерполяции нет совпадающих, т.е. если $x_i \neq x_j$ при $i \neq j$. Решив эту систему найдем коэффициенты интерполяционного многочлена (6.25). Заметим вместе с тем, что такой путь построения интерполяционного многочлена требует значительного объема вычислений, особенно при большом числе узлов. Существуют более простые алгоритмы построения интерполяционных многочленов.

Будем искать многочлен в виде линейной комбинации многочленов степени n:

$$L(x) = y_0 l_0(x) + y_1 l_1(x) + ... + y_n l_n(x) \tag{5.27}$$

При этом потребуем, чтобы каждый многочлен $l_i(x)$ обращался в 0 на всех узлах интерполяции, за исключением одного (i-го), где он должен равняться единице. Легко проверить, что этим условиям отвечает многочлен вида

$$l_0(x) = \frac{(x-x_1)(x-x_2)...(x-x_n)}{(x_0-x_1)(x_0-x_2)...(x_0-x_n)} \tag{5.28}$$

Действительно, $l_0(x_0)=1$ при $x=x_0$. При $x=x_1, x_2, \ldots, x_n$ числитель выражения (6.28) обращается в нуль. По аналогии с (6.28) получим

$$\begin{aligned} l_1(x) &= \frac{(x-x_0)(x-x_2)...(x-x_n)}{(x_1-x_0)(x_1-x_2)...(x_1-x_n)}, \\ l_2(x) &= \frac{(x-x_0)(x-x_1)(x-x_3)...(x-x_n)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)...(x_2-x_n)} \\ l_i(x) &= \frac{(x-x_1)...(x-x_{i-1})(x-x_{i+1})...(x-x_n)}{(x_i-x_0)...(x_i-x_{i-1})(x_i-x_{i+1})...(x_i-x_n)} \end{aligned} \tag{5.29}$$

Подставляя в (6.27) выражения (6.28), (6.29), находим

$$L(x) = \sum_{i=0}^{n} y_i \frac{(x-x_0)...(x-x_{i-1})(x-x_{i+1})...(x-x_n)}{(x_i-x_0)...(x_i-x_{i-1})(x_i-x_{i+1})...(x_i-x_n)} \tag{5.30}$$

Эта формула называется интерполяционным многочленом Лагранжа.

Покажем, что этот многочлен является единственным. Допустим противоположное: пусть существует еще один многочлен F(x) степени n, принимающий в узлах интерполяции заданные значения, т.е. $F(x_i)=y_i$. Тогда разность $R(x)=L(x)-F(x)$, являющийся многочленом степени n (или ниже), в узлах x_i равна

$R(x_i)=L(x_i)-F(x_i)=0$, i=0,1,…,n.

Это означает, что многочлен R(x) степени не больше n имеет n+1 корней. Отсюда следует, что R(x)=0 и L(x)=F(x).

Из формулы (6.30) можно получить выражение для линейной (n=1) и квадратичной (n=2) интерполяций:

$$L(x) = \frac{x-x_1}{x_0-x_1} y_0 + \frac{x-x_0}{x_1-x_0} y_1, \text{ n=1;}$$

$$L(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} y_0 + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} y_2, \text{ n=2.}$$

Существует несколько обобщений интерполяционного многочлена Лагранжа. Например, довольно широко используются интерполяционные многочлены Эрмита. Здесь наряду со значениями функции y_i'. Здесь задача состоит в том, чтобы найти многочлен $\varphi(x)$ степени 2n+1, значения которого и его производной в узлах x_i

удовлетворяют соответственно соотношениям

$$\varphi(x_i) = y_i,\ \varphi'(x_i) = y_i',\ i = 0,1,...,n.$$

В этом случае так же существует единственное решение, если все x_i различны.

3.3 Многочлен Ньютона

До сих не делалось никаких предположений о законе распределения узлов интерполяции. Теперь рассмотрим случай равноотстоящих значений аргумента, т.е. $x_i - x_{i-1} = h = \text{const}$ (i=1,2…,n). Величина h называется шагом.
Введем также понятие конечных разностей. Пусть известны значения функции в узлах x_i: $y_i = f(x_i)$.
Составим разности значений функции:

$$\Delta y_0 = y - y_0 = f(x_0 + h) - f(x_0),$$

$$\Delta y_1 = y_2 - y_1 = f(x_0 + 2h) - f(x_0 + h),$$

……..

$$\Delta y_{n-1} = y_n - y_{n-1} = f(x_0 + nh) - f(x_0 + (n-1)h).$$

Эти значения называют первыми разностями (или разностями первого порядка) функции.
Можно составить вторые разности функции:

$$\Delta^2 y_0 = \Delta y_1 - \Delta y_0,\ \Delta^2 y_1 = \Delta y_2 - \Delta y_1, \ldots$$

Аналогично составляются разности порядка k:

$$\Delta^k y_i = \Delta^{k-1} y_{i+1} - \Delta^{k-1} y_i,\ \text{i=0, 1, …,n-1.}$$

Конечные разности можно выразить непосредственно через значения функций. Например,

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

$$\Delta^3 y_0 = \Delta^2 y_1 - \Delta^2 y_0 = \ldots = y_3 - 3y_2 + 3y_1 - y_0.$$

Аналогично для любого k можно написать

$$\Delta^k y_0 = y_k - k y_{k-1} + \frac{k(k-1)}{2!} y_{k-2} + \ldots + (-1)^k y_0 \qquad (5.31)$$

Эту формула можно записать и для значения разности в узле x_i:

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

Используя конечные разности, можно определить y_k :

$$y_k = y_0 + k\Delta y_0 + \frac{k(k-1)}{2!}\Delta^2 y_0 + \ldots + \Delta^k y_0 \qquad (5.32)$$

Перейдем к построению интерполяционного многочлена. Ньютона. Этот многочлен будем искать в следующем виде:

$$N(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \ldots + a_n(x - x_0)(x - x_1)\ldots(x - x_{n-1}) \qquad (5.33)$$

График многочлена должен проходить через заданные узлы, т.е. $N(x_i) = y_i$ (i=0, 1,…,n). Эти условия используем для нахождения коэффициентов многочлена:

$$N(x_0) = a_0 = y_0,$$

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

$$N(x_2) = a_0 + a_1(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1) = a_0 + 2a_1 h + 2a_2 h^2 = y_2$$

Найдем отсюда коэффициенты a_0, a_1, a_2:

$$a_0 = y_0,\ a_1 = \frac{y_1 - a_0}{h} = \frac{y_1 - y_0}{h} = \frac{\Delta y_0}{h},$$

$$a_2 = \frac{y_2 - a_0 - 2a_1 h}{2h^2} = \frac{y_2 - y_0 - 2\Delta y_0}{2h^2} = \frac{\Delta^2 y_0}{2h^2}.$$

Аналогично можно найти и другие коэффициенты. Общая формула имеет вид

$$a_k = \frac{\Delta^h y_0}{k! h^k},\ \text{k=0, 1, …, n.}$$

Подставляя эти выражения в формулы (6.33), получаем следующий вид интерполяционного многочлена Ньютона:

$$N(x) = y_0 + \frac{\Delta y_0}{h}(x - x_0) + \frac{\Delta^2 y_0}{2!h^2}(x - x_0)(x - x_1) + ... + \frac{\Delta^n y_n}{n!h^n}(x - x_0)(x - x_1)...(x - x_{n-1}) \quad (5.34)$$

Конечные разности $\Delta^k y_0$ могут быть вычислены по формуле (6.31).

Интерполяционный многочлен Ньютона можно записать в виде

$$N(x_i + th) = y_i + t\Delta y_i + \frac{t(t-1)}{2!}\Delta^2 y_i + ... + \frac{t(t-1)...(t-n+1)}{n!}\Delta^n y_i \quad (5.35)$$

i=0,1,…

Полученное выражение называется первым интерполяционным многочленом Ньютона для интерполирования вперед.

Интерполяционную формула (6.35) обычно используют для вычисления значений функции в точках левой половины рассматриваемого отрезка. Это объясняется следующим. Разности $\Delta^k yi$ вычисляются через значения функций y_i, $y_{i+1},\ldots,y_{i+k}$ причем $i+k \le n$; поэтому при больших значениях i мы не можем вычислить разности высших порядков ($k \le n - i$). Например, при i=n-3 в (6.35) можно учесть только $\Delta y, \Delta^2 y, \Delta^3 y$.

Для первой половины рассматриваемого отрезка разности лучше вычислять справа налево. В этом случае

$$t=(x-x_n)/h \quad (5.36)$$

т.е. t<0, и интерполяционный многочлен можно получить в виде

$$N(x_n + th) = y_n + t\Delta y_{n-1} + \frac{t(t+1)}{2!}\Delta^2 y_{n-2} + ... + \frac{t(t+1)...(t+n-1)}{n!}\Delta^n y_0 \quad (5.37)$$

Полученная формула называется вторым интерполяционным многочленом Ньютона для интерполирования назад.

Рассмотрим пример применения интерполяционной формулы Ньютона при ручном счете.

Пример. Вычислить в точках x=0.1, 0.9 значения функции y=f(x), заданной таблице 6.1.

Процесс вычислений удобно свести в ту же Таблица 5.1. Каждая последующая конечная разность получается путем вычитания в предыдущей колонке верхней строки из нижней.

Таблица 5.1

X	Y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$	$\Delta^5 y$
0	1.2715	1.1937	-0.0146	0.0007	-0.0001	0.0000
0.2	2.4652	1.1791	—0.0139	0.0006	—0.0001	
0.4	3.6443	1.1652	—0.0133	0.0005		
0.6	4.8095	l.f519	—0.0128			
0.8	5.9614	1.1391				
1.0	7.1005					

При X=0.1 имеем $t=(x-x_0)/h=(0.1-0)/0.2=0.5$ По формуле получим

$$f(0.1) \approx N(0.1) = 1.2715 + 0.5 \cdot 1.1937 + \frac{0.5(0.5-1)}{2!} \cdot (-0.0146) +$$

$$+ \frac{0.5(0.5-1)(0.5-2)}{3!} \cdot 0.0007 + \frac{0.5(0.5-1)(0.5-2)(0.5-3)}{4!} \cdot (-0.0001) = 1.2715 + 0.59685 +$$

$$+ 0.00004 + 0.000004 = 1.8702.$$

Для сравнения по формуле линейной интерполяции получаем

$$f(0.1) \approx 1.8684.$$

Значение функции в точке x=0.9. $t=(x-x_n)/h=(0.9-1)/0.2=-0.5$. Тогда

$$f(0.9) \approx N(0.9) = 7.1005 - 0.5 \cdot 1.1391 - \frac{0.5(-0.5+1)}{2!} \cdot (-0.0128) - \frac{0.5(-0.5+1)(-0.5+2)}{3!} \cdot 0.0005 -$$

$$- \frac{0.5(-0.5+1)(-0.5+2)(-0.5+3)}{4!} \cdot (-0.0001) = 7.1005 - 0.5696 + 0.0016 - 0.00003 + 0.0000004 =$$

$$= 6.5325$$

Мы рассмотрели построение интерполяционного многочлена Ньютона для равноотстоящих узлов. Можно построить многочлен Ньютона и для произвольно расположенных узлов, как и в случае многочлена Лагранжа.

В заключении отметим, что разные способы построения многочленов Лагранжа и Ньютона дают тождественные интерполяционные формулы при заданной таблицу значений функции. Это следует из единственности интерполяционного многочлена заданной степени (при отсутствии совпадающих узлов интерполяции).

3.4 Сплайны

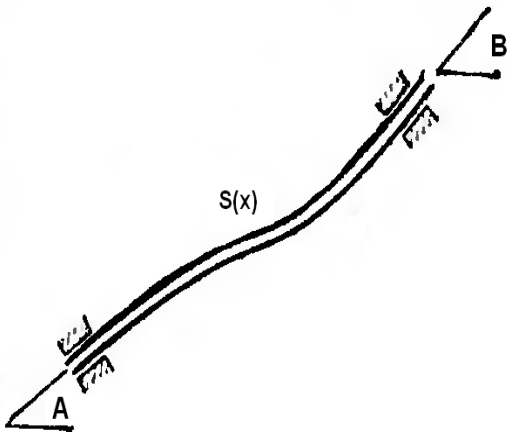

Рисунок 5.3

Сейчас широкое распространение для интерполяции получило использование кубических сплайн-функций — специальным образом построенных многочленов третьей степени. Они представляют собой некоторую математическую модель гибкого тонкого стержня из упругого материала. Если закрепить его в двух соседних узлах интерполяции с заданными углами наклонов A и B (Рисунок 5.3), то между точками закрепления этот стержень (механический сплайн) примет некоторую форму, минимизирующую его потенциальную энергию. Пусть форма этого стержня определяется функцией y=S(x). Из курса сопротивления материалов известно, что уравнение свободного равновесия имеет вид $S^{IV}(x)= 0$. Отсюда следует, что между каждой парой соседних узлов интерполяции функция S(x)является многочленом третьей степени. Запишем ее в виде

$$S(x) = a_i + b(x - x_{i-1}) + c(x - x_{i-1})^2 + d_i(x - x_{i-1})^3 \qquad (5.38)$$

$x_{i-1} \le x \le x_i$

Для определения коэффициентов a_i, b_i, c_i, d_i на всех n элементарных отрезках необходимо получить 4n уравнений. Часть из них вытекает из условий прохождения графика функции S(x) через заданные точки, т.е $S(x_{i-1})=y_{i-1}$. $S(x_i)=y_i$. Эти условия можно записать в виде

$$S(x_{i-1}) = a_i = y_{i-1} \qquad (5.39)$$

$$S(x_i) = a_i + b_i h_i + c_i h^2{}_i + d_i h_i^3 = y_i \qquad (5.40)$$

$h_i = x_i - x_{i-1}$, i=1, 2, , …, n.

Эта система содержит 2n уравнении. Для получения недостающих уравнений зададим условия непрерывности первых и вторых производных в узлах интерполяции, т. е. условия гладкости кривой во всех точках.

Вычислим производные многочлена (6.38).

$$S'(x) = b_i + 2c_i(x - x_{i-1}) + 3d_i(x - x_{i-1})^2,$$

$$S''(x) = 2c_i + 6d_i(x - x_{i-1}).$$

Приравнивая в каждом внутреннем узле $x = x_i$ значения этих производных, вычисленные в левом и правом от узла интервалах, получаем 2n—2 уравнений

$$b_{i+1} = b_i + 2c_i h_i + 3h_i^2 d_i \qquad (5.41)$$

$$c_{i+1} + c_i + 3h_i d_i \qquad (5.42)$$

i=1, 2, …, n-1.

Недостающие два соотношения получаются из условий закрепления концов сплайна.

В частности, при свободном закреплении концов можно приравнять нулю кривизну линии в этих точках. Такая функция, называемая свободным кубическим сплайном, обладает свойством минимальной кривизны, т. е. она самая гладкая среди всех интерполяционных функций данного класса. Из условий нулевой кривизны на концах следуют равенства нулю вторых производных в этих точках:

$$S''(x_0) = c_1 = 0, \qquad S''(x_n) = 2c_n + 6d_n h_n = 0 \tag{5.43}$$

Уравнения (6.39) – (6.43) составляют систему алгебраических уравнений для определения 4n коэффициентов a_i, b_i, c_i, d_i, (i=1,2, …,n).

Однако с целью аксиомой памяти ЭВМ и машинного времени эту систему можно привести к более удобному виду. Из условия (6.39) сразу можно найти все коэффициенты a_i. Далее из (6.42), (6.43) получим

$$d_i = \frac{c_{i+1} - c}{3h_i},\ \text{i=1,2,…,n-1},\ d_n = -\frac{c_n}{3h_n} \tag{5.44}$$

Подставим эти соотношения, а также значения $a_i=y_i$-1 в (6.40) и найдем отсюда коэффициенты

$$b_i = \frac{y_i - y_{i-1}}{h_i} - \frac{h_i}{3}(c_{i+1} + 2c_i),\ \text{i=1,2,…,n-1.}$$

$$b_n \frac{y_n - y_{n-1}}{h_n} - \frac{3}{2}h_n c_n \tag{5.45}$$

Учитывая выражения (6.44) и (6.45), исключаем из уравнения (6.40) коэффициенты d_i и b_i. Окончательно получим следующую систему уравнений только для коэффициентов c_i:

$$c_1 = 0, \qquad c_{n+1} = 0,$$

$$h_{i-1}c_{i-1} + 2(h_{i-1} + h_i)c_i + h_i c_{i+1} = 3\left(\frac{y_i - y_{i-1}}{h_i} - \frac{y_{i-1} - y_{i-2}}{h_{i-1}}\right) \tag{5.46}$$

i=1,2,…,n.

Матрица этой системы трехдиагональная, т. е. ненулевые элементы находятся лишь на главной и двух соседних с ней диагоналях, расположенных сверху и снизу. Для ее решения целесообразно использовать метод прогонки. По найденным из системы (6.46) коэффициентам c_i легко вычислить коэффициенты d_i и b_i.

3.5 Точность интерполяции

График интерполяционного многочлена y=F(x) проходит через заданные точки, т.е. значения многочлена и данной функции y=f(x) совпадают в узлах $x=x_i$ (i=0, 1, …,n). Если функция f(x) сама является многочленом степени n, то имеет место тождественное совпадение: f(x)=F(x). В общем случае в точках, отличных от узлов интерполяции $R(x)= f(x)-F(x) \neq 0$. Эта разность есть погрешность интерполяции и называется остаточным членом интерполяционной формулы. Оценим его значение.

Предположим, что заданные числа y_i являются значениями некоторой функции y=f(x) в точках $x=x_i$.Пусть эта функция непрерывна и имеет непрерывные производные до n+1-го порядка включительно. Можно показать, что в этом случае остаточный член интерполяционного многочлена Лагранжа имеет вид

$$R_L(x) = \frac{(x - x_0)(x - x_1)...(x - x_n)}{(n+1)!} f^{(n+1)}(x_*) \tag{5.47}$$

Здесь $f^{(n+1)}(x_*)$ - производная n+1 го порядка функции f(x) в некоторой точке $x = x_* \in [x_0, x_n]$. Если максимальное значение этой производной равно

$$\max_{x_0 \le x \le x_n} \left| f^{(n+1)}(x) \right| = M_{n+1}$$

То можно записать формулу для оценки остаточного члена:

$$\left|R_L(x)\right| \le \frac{(x - x_0)(x - x_1)...(x - x_n)}{(n+1)!} M_{n+1}.$$

Остаточный член интерполяционного многочлена Ньютона можно записать в виде

$$R_N(x) = \frac{t(t-1)...(t-n)}{(n+1)!} f^{(n+1)}(x_*)h^{n+1},\ t = \frac{x - x_0}{h}$$

Выбор способа интерполяции определяется различными соображениями: точностью, временем вычислении, погрешностями округлений и др. В некоторых случаях более предпочтительной может оказаться локальная интерполяция, в то время как построение единого многочлена высокой степени (глобальная интерполяция) не приводит к успеху.

Такого рода ситуацию в 1901 г. обнаружил К. Рунге. Он строил на отрезке $-1 \le x \le 1$ интерполяционные многочлены с равномерным распределением узлов для функции $y = 1/(1 + 25x^2)$. Оказалось, что при

увеличении степени интерполяционного многочлена последовательность его значении расходится для любой фиксированной точки x при $0.7 < |x| < 1$.
Положение в некоторых случаях может быть исправлено специальным распределением узлов интерполяции (если они не зафиксированы). Доказано, что если функция f(x) имеет непрерывную производную на отрезке [—1, 1], то при выборе значении x_i, совпадающих с корнями многочленов Чебышева степени n + 1, интерполяционные многочлены степени n сходятся к значениям функции в любой точке этого отрезка.
Таким образом, повышение точности интерполяции целесообразно производить за счет уменьшения шага и специального расположения точек x_i. Повышение степени интерполяционного многочлена при локальной интерполяции также уменьшает погрешность, однако здесь не всегда ясно поведение производной $f^{(n+1)}(x_*)$ при увеличении n. Поэтому на практике стараются использовать многочлены малой степени (линейную и квадратичную интерполяции, сплайны).

§4.Аппроксимация. Метод наименьших квадратов. Эмпирические формулы. Локальное сглаживание данных.

4.1 Аппроксимация методом наименьших квадратов (МНК)

Метод наименьших квадратов (часто называемый МНК) обычно упоминается в двух контекстах. Во-первых, широко известно его применение в регрессионном анализе, как метода построения моделей на основе зашумленных экспериментальных данных. При этом помимо собственно построения модели обычно осуществляется оценка погрешности, с которой были вычислены её параметры, иногда решаются и некоторые другие задачи. Во-вторых, МНК часто применяется просто как метод аппроксимации, без какой-либо привязки к статистике. На этой странице МНК рассматривается как метод аппроксимации.
Общий линейный метод наименьших квадратов
При аппроксимации методом наименьших квадратов аппроксимируемая функция f задается набором N точек (x_i , y_i). Аппроксимирующая функция g строится, как линейная комбинация базисных функций Fj (количество функций M обычно меньше числа точек N)

$$g = \sum_{j=0}^{M-1} c_j F_j(x)$$

При этом коэффициенты c_j выбираются таким образом, чтобы минимизировать сумму квадратов отклонений аппроксимирующей функции от заданных значений

$$E = \sum_{i=0}^{N-1} \left(y_i - \sum_{j=0}^{M-1} c_j F_j(x_i) \right)^2$$

Иногда, если различным точкам назначен различный вес wi , каждое слагаемое в сумме квадратов умножается на квадрат соответствующего ему веса:

$$E = \sum_{i=0}^{N-1} w_i^2 \left(y_i - \sum_{j=0}^{M-1} c_j F_j(x_i) \right)^2$$

Такая задача называется построением взвешенной аппроксимации по МНК. Следует отметить, что выбор именно такого способа аппроксимации (линейной комбинации базисных функций) и именно такой оценочной функции (суммы квадратов отклонений) является далеко не единственно возможным. Базисные функции могут входить в функцию g нелинейно, а оценочная функция E может быть заменена максимумом отклонения или любой другой функцией оценки. Однако именно такой способ аппроксимации с такой оценочной функцией позволяет нам найти наилучшие cj за конечное число операций, сведя задачу к решению системы линейных уравнений.
Методы поиска коэффициентов
Перед тем, как рассматривать способы поиска коэффициентов cj , введем следующие обозначения:

$$A = \begin{pmatrix} a_{11} & a_{12} & .. \\ a_{21} & a_{22} & .. \\ .. & .. & .. \\ a_{n1} & .. & a_{nm} \end{pmatrix}, \quad a_{ij} = w_i F_j(x_i)$$

$$b = \begin{pmatrix} b_1 \\ .. \\ .. \\ b_n \end{pmatrix}, \quad b_i = w_i y_i$$

$$c = \begin{pmatrix} c_1 \\ .. \\ .. \\ c_n \end{pmatrix}$$

Первый способ поиска оптимальных cj состоит в решении системы уравнений $\frac{\partial E}{\partial c_j} = 0$(т.н. метод нормальных уравнений). С учетом особенностей задачи, равенство производных нолю является не только необходимым, но и достаточным условием минимума. Кроме того, эта система уравнений является линейной относительно c_j и записывается как !!!!(A TA)c - A Tb = 0. Основным недостатком этого метода является то, что полученная система уравнений часто бывает плохо обусловленной. Число обусловленности матрицы A само по себе может быть велико (например, в случае использования базиса из полиномов), а возведение матрицы A в квадрат увеличивает его ещё больше (также возводит в квадрат). По этой причине метод нормальных уравнений на практике обычно не применяется.

Второй способ основан на том, что задачу минимизации функции E можно записать в матричной форме, как поиск $\min_c \|Ac - b\|_2$. Таким образом, задача сведена к поиску псевдорешения системы линейных уравнений, которое может быть найдено с использованием сингулярного разложения (SVD). Этот способ является наиболее естественным, хотя и требует более сложного программного кода для своей реализации.

Достоинством этого метода является то, что в этом случае не происходит возведения в квадрат числа обусловленности базиса. Мы решаем систему, обусловленность которой равна обусловленности базиса, что позволяет получать достаточно точные решения даже в тех случаях, когда использование метода нормальных уравнений приводит к возникновению системы с вырожденной матрицей коэффициентов. Следует отметить, что хотя для осуществления сингулярного разложения мог бы использоваться предназначенный для этого общий алгоритм, в этом нет необходимости. С учетом специфики задачи предпочтительнее использовать более компактную и быструю версию алгоритма, учитывающую тот факт, что требуется не само сингулярное разложение матрицы A, а произведение отдельных компонент этого разложения на вектор b.

Также существует третий способ, применяющий QR-разложение, и работающий несколько быстрее метода на основе SVD-разложения. Сначала матрица A представляется в виде произведения прямоугольной ортогональной матрицы Q и квадратной верхнетреугольной матрицы R. Затем решается система уравнений Rx = Q Tb. Достоинством этого метода является относительно высокая скорость работы, недостатком - применимость только к невырожденным задачам (т.е. задачам, имеющим одно и только одно решение). Для того, чтобы задача аппроксимации была невырожденной, требуется выполнение двух условий. Во-первых, система базисных функций должна быть линейно-независимой. Во-вторых, число точек должно быть не меньше числа базисных функций. Как правило, в практических задачах эти условия выполняются. Но все же очень редко встречаются задачи, в которых алгоритм на основе QR-разложения оказывается неприменим из-за нарушения одного из условий. Кроме того, обычно число строк в матрице A намного больше числа столбцов, а правильно реализованное SVD разложение таких матриц по трудоемкости сравнимо с QR разложением. Таким образом, более целесообразным представляется использование второго способа, т.е. SVD-разложения.

4.2 Линейная аппроксимация по МНК

При линейной аппроксимации по МНК в качестве набора базисных функций используются f0 = 1 и f1 = x. Линейная комбинация функций из этого базиса позволяет получить произвольную прямую на плоскости.

Особенностью этого базиса является то, что в данном случае оказывается применим метод нормальных уравнений. Базис небольшой, хорошо обусловленный, и поэтому задача аппроксимации легко сводится к решению системы линейных уравнений 2x2. Для решения используется модификация метода вращений, которая позволяет корректно обрабатывать вырожденные случаи (например, точки с совпадающими абсциссами).

Этот алгоритм реализован в подпрограмме BuildLinearLeastSquares.

Полиномиальная аппроксимация по МНК

Частой ошибкой является использование для полиномиальной аппроксимации следующего набора базисных функций: fi = x_i. Этот базис является наиболее естественным и, пожалуй, первым приходящим в голову вариантом. Но хотя с этим базисом удобно работать, он очень плохо обусловлен. Причем проблемы появляются даже при умеренном числе базисных функций (порядка десяти). Эта проблема была решена путем выбора базиса из полиномов Чебышева вместо базиса из степеней x. Полиномы Чебышева линейно выражаются через степени x и наоборот, так что эти базисы эквивалентны. Однако обусловленность базиса из полиномов Чебышева значительно лучше, чем у базиса из степеней x. Это позволяет без проблем осуществлять аппроксимацию полиномами практически любой степени.

Алгоритм работает следующим образом. Входной набор точек масштабируется и приводится к интервалу [−1,1], после чего на этом отрезке строится базис из полиномов Чебышева. После аппроксимации по МНК пользователь получает набор коэффициентов при полиномах Чебышева. Эта задача решается подпрограммой BuildChebyshevLeastSquares. Полученные коэффициенты могут быть переданы в подпрограмму CalculateChebyshevLeastSquares, которая вычисляет значение аппроксимирующей функции в требующей точке.

Если специфика решаемой задачи требует использования именно базиса из степеней x, то можно воспользоваться подпрограммой BuildPolynomialLeastSquares. Эта подпрограмма решает задачу с

использованием полиномов Чебышева, после чего преобразовывает результат к степеням x. Это позволяет отчасти сохранить точность вычислений, т.к. все промежуточные расчеты ведутся в хорошо обусловленном базисе. Однако все же предпочтительным вариантом является базис из полиномов Чебышева.
В некоторых случаях требуется решение задачи с ограничениями. Например, мы можем знать, что аппроксимируемая функция равна 1 на границе, т.е. при x = 0. Если у нас имеется достаточно большое количество точек в окрестностях x = 0, то построенная функция пройдет в окрестностях точки (0,1). Однако точного совпадения скорее всего не будет. Во многих случаях точное соблюдение каких-то граничных условий не требуется. Однако иногда встречаются задачи, требующие именно точного прохождения аппроксимирующей функции через несколько заданных точек или равенства производной в некоторых точках заранее заданным значениям. При этом особые требования предъявляются только к нескольким точкам xk . В остальных же точках мы просто минимизируем сумму квадратов отклонений.
В таком случае имеет место задача аппроксимации с ограничениями вида

$$g(x_k) = g_k$$

или

$$\frac{dg(x_k)}{dx} = \tilde{g}_k$$

Для решения таких задач предназначен специальный алгоритм, реализованный в подпрограмме BuildChebyshevLeastSquaresConstrained. В настоящее время подпрограмма способна обрабатывать произвольное количество ограничений (но строго меньшее, чем M) на значение функции и её первой производной. Ограничения на значения производных высших порядков не поддерживаются.

4.3Аппроксимация кубическими сплайнами

Ещё одним возможным набором базисных функций являются кубические сплайны. Множество сплайнов с общими узлами образует линейное пространство, что позволяет применить к ним линейный метод наименьших квадратов. В качестве базисных функций выбираются сплайны, удовлетворяющие следующим условиям:

$$S_j(i) = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases}$$

$$i, j = 0 \ldots M-1$$

Для простоты мы предполагаем, что аппроксимация строится на отрезке [0,M-1], с узлами, равномерно распределенными по отрезку. Разумеется, в реальности отрезок, на котором осуществляется аппроксимация, может быть любым (однако равномерность распределения узлов сохраняется). На графике ниже приведен пример такой системы базисных функций:

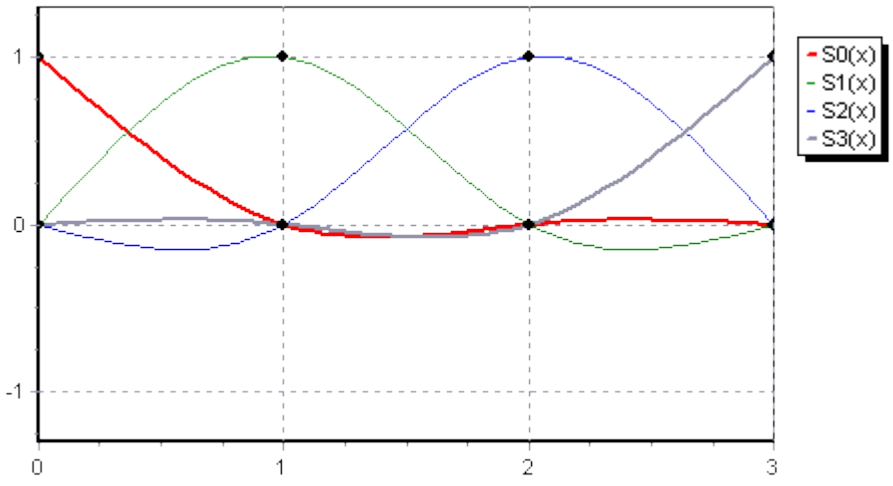

Рисунок 5.4

Построение аппроксимирующей функции осуществляется подпрограммой BuildSplineLeastSquares. Результатом её работы является массив коэффициентов, задающий результирующий кубический сплайн. Этот массив передается в подпрограмму SplineInterpolation (см. описание модуля для интерполяции кубическими сплайнами), которая рассчитывает значение аппроксимирующего сплайна в указанной точке.

4.4 Аппроксимация произвольным набором базисных функций

Выше были приведены три набора базисных функций, которые могут использоваться для аппроксимации: прямые, полиномы и кубические сплайны. Однако это не единственные наборы функций, которые могут быть

применены. И, разумеется, нельзя предусмотреть все возможные практические ситуации и то, какой набор функций окажется востребован. Поэтому любой программный пакет, решающий задачу аппроксимации, должен содержать в себе подпрограмму для аппроксимации произвольным набором базисных функций. Эту задачу решает подпрограмма BuildGeneralLeastSquares.
Поскольку заранее неизвестно, какой именно набор функций используется, в подпрограмму требуется передать информацию о значениях базисных функций в точках xi . Эта информация передается в виде матрицы FMatrix, содержащей значение i-ой базисной функции в j-ой точке. Также в подпрограмму передаются массивы y и w, содержащие ординаты исходных точек и их веса.
Обратите внимание, что абсциссы xi в подпрограмму не передаются. Дело в том, что при аппроксимации общим МНК абсциссы точек участвуют в процессе только как аргументы базисных функций, а эта информация уже содержится в матрице FMatrix. То есть процесс аппроксимации не зависит от размерности пространства, в котором она строится - достаточно пронумеровать точки, вычислить в них значения функции и передать информацию в алгоритм. Это позволяет использовать подпрограмму как для решения одномерных задач, так и для многомерных проблем.

4.5 Подбор эмпирических формул

Характер опытных данных.
При интерполировании функции мы использовали условие равенства значений интерполяционного многочлена и данной функции в известных точках — узлах интерполяции. Это предъявляет высокие требования к точности данных значений функции. В случае обработки опытных данных, полученных в результате наблюдении или измерении, нужно иметь в виду ошибки этих данных. Они могут быть вызваны несовершенством измерительного прибора, субъективными причинами, различными случайными факторами и т. д. Ошибки экспериментальных данных можно условно разбить на три категории по их происхождению и величине: систематические, случайные и грубые.
Систематические ошибки обычно дают отклонение в одну сторону от истинного значения измеряемой величины. Они могут быть постоянными пли закономерно изменяться при повторении опыта, и их причина и характер известны. Систематические ошибки могут быть вызваны условиями эксперимента (влажностью, температурой среды и др.), дефектом измерительного прибора, его плохой регулировкой (например, смещением указательной стрелки от нулевого положения) и т. д. Эти ошибки можно устранить наладкой аппаратуры или введением соответствующих поправок.
Случайные ошибки определяются большим числом факторов, которые не могут быть устранены либо достаточно точно учтены при измерениях или при обработке результатов. Они имеют случайный (несистематический) характер, дают отклонении от средней величины в ту и другую стороны при повторении измерении и не могут быть устранены в эксперименте, как бы тщательно он ни проводился. С вероятностной точки зрения математическое ожидание случайной ошибки равно нулю. Статистическая обработка экспериментальных данных позволяет определить величину случайной ошибки и довести ее до некоторого приемлемого значения повторением измерений достаточное число раз.
Грубые ошибки явно искажают результат измерения; они чрезмерно большие и обычно пропадают при повторении опыта. Грубые ошибки существенно выходят за пределы случайной ошибки, полученные при статистической обработке. Измерения с такими ошибками отбрасываются и в расчет при окончательной обработке результатов измерений не принимаются.
Таким образом, в экспериментальных данных всегда имеются случайные ошибки. Они, вообще говоря, могут быть уменьшены до сколь угодно малой величины путем многократного повторения опыта. Однако это не всегда целесообразно, поскольку могут потребоваться большие материальные пли временные ресурсы. Значительно дешевле и быстрее можно в ряде случаев получить уточненные данные хорошей математической обработкой имеющихся результатов измерений.
В частности, с помощью статистической обработки результатов измерений можно найти закон распределения ошибок измерений, наиболее вероятный диапазон изменения искомой величины (доверительный интервал) и другие параметры. Рассмотрение этих вопросов выходит за рамки данного пособия; их изложение можно найти в некоторых книгах, приведенных в списке литературы. Здесь ограничимся лишь определением связи между исходным параметром x и искомой величиной y на основании результатов измерений.

4.6 Эмпирические формулы.

Пусть, изучая неизвестную функциональную зависимость между y и x, мы в результате серии экспериментов произвели ряд измерений этих величин и получили таблицу значении

x_0	x_1	$\cdots$	x_n
y_0	y_1	$\cdots$	y_n

Задача состоит в том, чтобы найти приближенную зависимость

$$y = f(x) \tag{5.48}$$

значения которой при $x = x_i \quad (i = 0,1,...,n)$ мало отличаются от опытных данных y_i. Приближенная

функциональная зависимость (6.48), полученная на основании экспериментальных данных, называется *эмпирической формулой.*

Задача построения эмпирической формулы отличается от задачи интерполирования. График эмпирической зависимости, вообще говоря, не проходят через заданные точки (x_i, y_i), как в случае интерполяции. Это приводит к тому, что экспериментальные данные в некоторой степени сглаживаются, а интерполяционная формула повторила бы все ошибки, имеющиеся в экспериментальных данных.

Построение эмпирической формулы состоит из двух этапов: подбора общего вида этой формулы и определения наилучших значений содержащихся в ней параметров. Общий вид формулы иногда известен из физических соображений. Например, для упругой среды связь между напряжением σ и относительной деформацией ε определяется законом Гука: $\sigma = E\varepsilon$, где E — модуль упругости; задача сводится к определению одного неизвестного параметра E.

Если характер зависимости неизвестен, то вид эмпирической формулы может быть произвольным. Предпочтение обычно отдается наиболее простым формулам, обладающим достаточной точностью. Они первоначально выбираются из геометрических соображений: экспериментальные точки наносятся на график и примерно угадывается общий вид зависимости путем сравнения полученной кривой с графиками известных функций (многочлена, показательной пли логарифмической функций и т. п.). Успех здесь в значительной мере определяется опытом и интуицией исследователя.

Простейшей эмпирической формулой является линейная зависимость

$$y = ax + b \tag{5.49}$$

Близость экспериментального распределения точек к линейной зависимости легко просматривается после построения графика данной экспериментальной зависимости, Кроме того, эту зависимость можно проверить путем вычисления значений k_i:

$$k_i = \Delta y_i / \Delta x_i, \ \Delta y_i = y_{i+1} - y_i, \ \Delta x_i = x_{i+1} - x_i, \ i = 0,1,...,n-1.$$

Если при этом $k_i \approx const$, то точки (x_i, y_i) расположены приблизительно на одной прямой, и может быть поставлен вопрос о применимости эмпирической формулы (6.49). Точность такой аппроксимации определяется отклонением величин k_i от постоянного значения. В частном случае равноотстоящих точек x_i, (т. е. $\Delta x_i = const$) достаточно проверить постоянство разностей Δy_i.

Пример. Проверим возможность использования линейной зависимости для описания следующих данных:

Поскольку здесь x_i — равноотстоящие точки ($\Delta x_i = x_{i+1} - x_i = 0.5$), то достаточно вычислить разности Δy_i: 0.64, 0.69, 0.65, 0.64, 0.65. Так как эти значения близки друг к другу, то в качестве эмпирической формулы можно принять линейную зависимость.

В ряде случаев к линейной зависимости могут быть сведены и другие экспериментальные данные, когда их график в декартовой системе координат не является прямой линией. Это может быть достигнуто путем введения новых переменных ξ, η вместо x и y:

$$\xi = \varphi(x, y), \qquad \eta = \psi(x, y) \tag{5.50}$$

Функции $\varphi(x, y)$ и $\psi(x, y)$ выбираются такими, чтобы точки (ξ_i, η_i) лежали на некоторой прямой линии в плоскости (ξ, η). Такое преобразование называется *выравниванием данных.*

Для получения линейной зависимости

$$\eta = a\xi + b$$

с помощью преобразования (6.50) исходная формула должна быть записана в виде

$$\psi(x, y) = a\varphi(x, y) + b.$$

К такому виду легко сводится, например, степенная зависимость $y = ax^b$ (x>0, y>0). Логарифмируя эту формулу, получаем $\lg y = b \lg x + \lg a.$ Полагая $\xi = \lg x$, $\eta = \lg y$, находим линейную связь; $\eta = b\xi + c$ $(c = \lg a).$

4.7 Локальное сглаживание данных.

Как отмечалось ранее , опытные данные содержат случайные ошибки, что является причиной разброса этих данных. Во многих случаях бывает целесообразно провести их сглаживание для получения более плавного характера исследуемой зависимости. Существуют различные способы сглаживания. Рассмотрим один из них, основанный на методе наименьших квадратов.

Пусть в результате экспериментального исследования зависимости y = f{x) получена таблица значений искомой функции $y_0, y_1,\ldots,y_n$ в точках $x_1, x_0,\ldots, x_n$. Значения аргумента x (предполагаются равноотстоящими, а опытные данные y_i — имеющими одинаковую точность. Предполагается также, что функция y = f(x) на произвольной части отрезка $[x_0,x_n]$ может быть достаточно хорошо аппроксимирована многочленом

некоторой степени n.

Рассматриваемый способ сглаживания состоит в следующем. Для нахождения сглаженного значения $\overline{y}_i$ в точке x_i выбираем по обе стороны от нее k значений аргумента. По опытным значениям рассматриваемой функции в этих точках $y_{i-h/2},...,y_{i-1},y_i,y_{i+1},...,y_{i+h/2}$ строим многочлен степени m с помощью метода наименьших квадратов (при этом $m \le k$). Значение полученного многочлена $\overline{y}_i$ в точке x_i и будет искомым (сглаженным) значением. Процесс повторяется для всех внутренних точек. Сглаживание значений, расположенных вблизи концов отрезка $[x_0, x_n]$, производится с помощью крайних точек.

Опыт показывает, что сглаженные значения $\overline{y}_i$, как правило, с достаточной степенью точности близки к истинным значениям. Иногда сглаживание повторяют. Однако это может привести к существенному искажению истинного характера рассматриваемой функциональной зависимости.

Приведем в заключение несколько формул для вычисления сглаженных опытных данных при различных m, k:

m=1:

$$\overline{y}_i = \frac{1}{3}(y_{i-1}+y_i + y_{i+1}), \quad \text{k=2,}$$

$$\overline{y}_i = \frac{1}{5}(y_{i-2} + y_{i-1} + y_i + y_{i+1} + y_{i+2}), \quad \text{k=4,}$$

$$\overline{y}_i = \frac{1}{7}(y_{i-3} + y_{i-2} + y_{i-1} + y_i + y_{i+1} + y_{i+2} + y_{i+3}), \quad \text{k=6,}$$

m=3:

$$\overline{y}_i = \frac{1}{35}(-3y_{i-2} + 12y_{i-1} + 17y_i + 12y_{i+1} - 3y_{i+2}), \quad \text{k=4,}$$

$$\overline{y}_i = \frac{1}{21}(-2y_{i-3} + 3y_{i-2} + 6y_{i-1} + 7y_i + 6y_{i+1} + 3y_{i+2} - 2y_{i+3}), \quad \text{k=6,}$$

$$\overline{y}_i = \frac{1}{231}(-21y_{i-4} + 14y_{i-3} + 39y_{i-2} + 54y_{i-1} + 59y_i + 54y_{i+1} + 39y_{i+2} + 14y_{i+3} - 21y_{i+4}), \quad \text{k=8;}$$

$$\overline{y}_i = \frac{1}{231}(-5y_{i-3} - 30y_{i-2} + 75y_{i-1} + 131y_i + 75y_{i+1} - 30y_{i+2} + 5y_{i+3}), \quad \text{k=5,}$$

$$\overline{y}_i = \frac{1}{429}(15y_{i-4} - 55y_{i-3} + 30y_{i-2} + 135y_{i-1} + 179y_i + 135y_{i+1} + 30y_{i+2} - 55y_{i+3} + 15y_{i+4}), \quad \text{k=8.}$$

В 2009 году Университет стал победителем многоэтапного конкурса, в результате которого определены 12 ведущих университетов России, которым присвоена категория «Национальный исследовательский университет». Министерством образования и науки Российской Федерации была утверждена программа его развития на 2009–2018 годы. В 2011 году Университет получил наименование «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики»

КАФЕДРА ИНФОРМАТИКИ И ПРИКЛАДНОЙ МАТЕМАТИКИ

Кафедра образована в 1976 году из сотрудников кафедры Вычислительной техники для подготовки специалистов в области программирования и методов вычислений. Кафедру возглавляет д.т.н., профессор О. Ф. Немолочнов, работающий в области создания систем автоматизации проектирования ЭВМ.
Долгое время основным научным направлением кафедры было создание систем автоматизированного контроля цифровой аппаратуры. Кафедра активно сотрудничала с такими организациями, как НИЦЭВТ в Москве, НПО Электроавтоматика в Ленинграде, КБЭ в Харькове, завод САМ в Минске.
Кафедра принимала активное участие в деятельности ФПК преподавателей при ЛИТМО, и некоторые курсы, созданные сотрудниками кафедры, тиражировались на всю страну.
В последнее десятилетие на кафедре продолжаются работы по исследованию методов построения контролирующих и диагностических тестов, по автоматизации проектно-конструкторских работ в оптике, по информационному обеспечению САПР. Сотрудники кафедры разрабатывают новые учебные программы и циклы лабораторных работ, ориентированные на привитие студентам всех специальностей навыков практической работы на ЭВМ. На кафедре ведутся работы по внедрению методов дистанционного обучения, создаются учебники и учебно-методические комплексы по различным дисциплинам, связанным с разработкой программного обеспечения.

Денисова Эльвира Викторовна
Кучер Алексей Владимирович

Краткий курс вычислительной математики

Учебное пособие

В авторской редакции
Редакционно-издательский отдел НИУ ИТМО
Зав. РИО Н.Ф. Гусарова
Лицензия ИД № 00408 от 05.11.99
Подписано к печати
Заказ №
Тираж
Отпечатано на ризографе

Редакционно-издательский отдел
Санкт-Петербургского национального
исследовательского университета
информационных технологий, механики
и оптики
197101, Санкт-Петербург, Кронверкский пр., 49