Министерство образования
и науки Российской Федерации

Уральский
федеральный
университет
имени первого Президента
России Б.Н.Ельцина

В.Г. Пименов

ЧИСЛЕННЫЕ МЕТОДЫ

Учебное электронное текстовое издание

Пособие предназначено для учебно-методического обеспечения унифицированных модулей бакалавриата всех направлений подготовки института математики и компьютерных наук.
Подготовлено: кафедрой вычислительной математики ИМКН.

Екатеринбург

2012

Пособие содержит лекции первой части двухсеместрового курса "Численные методы", который читается в институте математики и компьютерных наук кафедрой вычислительной математики для всех направлений подготовки бакалавриата. В пособие вошли разделы "Теория погрешностей", "Численные методы решения нелинейных уравнений", "Численные методы решения линейных и нелинейных систем", "Теория интерполяции", "Численное дифференцирование", "Численное интегрирование".

ОГЛАВЛЕНИЕ

1. ВВЕДЕНИЕ

Рассмотрим место изучаемой дисциплины в ряду предметов, связанных с математическим моделированием.

Предположим, что имеется объект любой природы. Его требуется изучить, выявив закономерности и характеристики. С помощью конкретной науки, относящейся к объекту, составляется математическая модель, в которой отражаются самые существенные характеристики объекта. Модель содержит связи между характеристиками, которые, как правило, выражаются математическими уравнениями. С целью установить закономерности необходимо исследовать математическую модель, например решить уравнения. Но чем адекватнее составлена математическая модель, тем она сложнее, и напрямую аналитические методы решения зачастую неприменимы. Кроме аналитических методов в настоящее время все активнее применяют численные методы, алгоритмы, основанные на аппроксимации исходной математической модели другими, реализуемыми в виде вычислительных операций, зачастую громоздких. Следует отметить,что для правильной картины изучения математической модели, численные методы должны быть дополнены качественными методами, которые составляют значительную часть математики. Громоздкость численных расчетов приводит к необходимости использования вычислительных устройств (в широком смысле этого слова), и именно развитие компьютеров и их программного обеспечения привело к большинству современных достижений в науке и технике. Следуя словам академика А.А.Самарского отметим сказанное тезисом: основу математического моделирования составляет триада: модель – алгоритм – программа.

В результате на выходе получаем некоторую характеристику изучаемого объекта, которую условно будем считать числом. Но эта характери-

стика в результате обработки каждым элементом триады в той или иной степени огрубляется, вносится погрешность. Задача состоит в её изучении на каждой стадии.

$A_{неустр.}$ - погрешность, которая вносится на стадии математического моделирования, называется *неустранимой погрешностью.*

Пример: Физический маятник, возможно, с трением, описывается дифференциальным уравнением второго порядка

$$\begin{cases} \ddot{\phi} + a\sin\phi + b\dot{\phi} = f(t), \\ \phi(0) = \alpha, \\ \dot{\phi}(0) = \beta; \end{cases}$$

где ϕ - угол отклонения маятника от положения равновесия.

Начальные условия α, β, а также коэффициенты уравнения a и b должны быть измерены некоторыми приборами, которые имеют определенную точность, которая и дает неустранимую погрешность.

$A_{метода}$ - погрешность, которая вносится на стадии составления численного алгоритма, называется *погрешностью метода.*

Например, приведенное уравнение физического маятника аналитически не решается. Можно, конечно, заменить $\sin\phi$ на ϕ и получим линейное уравнение второго порядка (математический маятник), решение которого выписывается в элементарных функциях, но такое упрощение модели можно сделать только при малых ϕ. Между тем, в теории численных методов разработаны эффективные методы решения дифференциальных уравнений, подобных уравнению физического маятника. К числу них относятся метод Эйлера, методы Рунге-Кутты, методы Адамса и другие. Подбирая методы и их параметры, прежде всего шаг дискретизации, можно получить решение с требуемой погрешностью метода. Изучение разнообразия

таких методов и способов оценки их погрешностей составляет основное содержание курса.

$A_{вычисл.}$ - погрешность, которую вносит вычислительное устройство, называется *вычислительной погрешностью.*

Так подавляющее большинство численных алгоритмов, в том числе и упомянутые методы, требует громадных вычислений для достижения хорошей точности, поэтому необходимо эффективно работающее вычислительное устройство. Но компьютер представляет информацию дискретно, поэтому любые действия с вещественными числами, наиболее распространенными в моделировании, вносят вычислительную погрешность. По этой причине число операций нужно по возможности оптимизировать, оптимизировать нужно также и вычислительные средства.

Все сказанное можно изобразить в виде схемы

$$\begin{array}{ccccccc} & A_{неуст.} & & A_{метода} & & A_{выч.} & \\ & \downarrow & & \downarrow & & \downarrow & \\ \text{обьект} \rightarrow & \boxed{модель} & \rightarrow & \boxed{метод} & \rightarrow & \boxed{выч.устройство} & \rightarrow \text{число} \end{array}$$

Формула полной погрешности

$$A_{пол.} = A_{неустр.} + A_{мет.} + A_{выч.}$$

Считается, что идеальная ситуация наблюдается тогда, когда все три вида погрешности примерно равны или, по крайней мере, имеют один порядок. Например, нет смысла применять очень точные алгоритмы и считать на вычислительном устройстве с большой степенью точности, если имеется большая неустранимая погрешность.

Основной задачей данной дисциплины является средняя часть триады – изучение численных алгоритмов и погрешностей метода. Однако, при

этом, обязательно нужно уметь учитывать влияние неустранимой погрешности, а также нужно уметь выбирать третью часть триады (программу) с соответствующим влиянием вычислительной погрешности на конечный результат.

2. ТЕОРИЯ ПОГРЕШНОСТЕЙ

2.1. Числа и характеристики их точности

По умолчанию будем считать,что числа рассматриваются вещественные.

Пусть x - идеальное вещественное число, а x^* - его приближение. *Абсолютной погрешностью* числа x будет являться величина A_{x^*}, удовлетворяющая соотношению $|x - x^*| \leqslant A_{x^*}$.

Относительной погрешностью числа x называется величина $\Delta_{x^*} = \frac{A_{x^*}}{|x^*|}$ $(|x^*| \neq 0)$.

Рассмотрим число в десятичном позиционном представлении. Цифра, начиная с первой ненулевой, называется *значащей*.

Например, 20.02 - все цифры значащие, 002.02 - третья цифра является значащей.

Цифра называется *верной*, если абсолютная погрешность числа не превосходит половины единицы соответствующего разряда. Все остальные цифры *сомнительные* (то есть $A_{x^*} \leqslant 0.5 \cdot 10^{-m}$)

Например,

1) число 21.145, и пусть $A_{x^*} = 0.01$. Тогда 211 - верные цифры, 45 - сомнительные.

2) число 3.1415926, $A_{x^*} = 0.007$. Верные цифры 31, в самом деле, посмотрим на 3.14, единица разряда последней цифры 1/100, ее половина 5/1000, а погрешность больше $5/1000 < 0.007$, следовательно 4 – цифра сомнительная при данной оценке погрешности.

Понятие верной цифры после запятой связано с понятием абсолютной погрешности, а понятие верной значащей с относительной погрешностью.

2.2. Погрешность арифметических действий

Сложение и вычитание.

Пусть точные значения некоторых чисел – $x_1, x_2, \ldots, x_n$, а их приближения $x_1^*, x_2^*, \ldots, x_n^*$. Известны погрешности $A_{x_1^*}, A_{x_2^*}, \ldots, A_{x_n^*}$. Пусть требуется сложить числа, т.е. посчитать $y = \sum\limits_{i=1}^{n} x_i$. В качестве приближения суммы возьмем $y^* = \sum\limits_{i=1}^{n} x_i^*$. Оценим погрешность

$$|y - y^*| = |\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i^*| \leqslant \sum_{i=1}^{n} |x_i - x_i^*| \leqslant \sum_{i=1}^{n} A_{x_i^*} = A_{y^*}$$

Получили утверждение

Теорема 1. *При сложении и вычитании абсолютные погрешности складываются.*

Рассмотрим относительную погрешность сложения.

$$\Delta_{y^*} = \frac{A_{y^*}}{|y^*|} = \frac{\sum\limits_{i=1}^{n} A_{x_i^*}}{|\sum\limits_{i=1}^{n} x_i^*|} = \frac{\sum\limits_{i=1}^{n} \Delta_{x_i^*} |x_i^*|}{|\sum\limits_{i=1}^{n} x_i^*|}$$

Предположим, что складываются числа одного знака, тогда

$$\Delta_{y^*} = \frac{\sum\limits_{i=1}^{n} \Delta_{x_i^*} |x_i^*|}{\sum\limits_{i=1}^{n} |x_i^*|} \leqslant \frac{\max \Delta_{x_i^*} \sum\limits_{i=1}^{n} |x_i^*|}{\sum\limits_{i=1}^{n} |x_i^*|} = \max \Delta_{x_i^*}$$

Аналогичным образом доказывается, что при сложений чисел одного знака выполняется $\min \Delta_{x_i^*} \leq \Delta_{y^*}$.

Однако при сложении чисел разных знаков (при вычитании) может наблюдаться эффект возрастания относительной погрешности.

2.3. Эффект вычитания близких чисел

Пример: Требуется посчитать

$$\sqrt{11} - \sqrt{10}$$

если

$$\sqrt{11} \approx 3,32;\ \sqrt{10} \approx 3,16$$

(все цифры верные).

Тогда

$$\sqrt{11} - \sqrt{10} \approx 0,16.$$

Оценим погрешность. Обозначим $x_1 = \sqrt{11}$, $x_2 = \sqrt{10}$, $x_1^* = 3,32$, $x_2^* = 3,16$, $y = \sqrt{11} - \sqrt{10}$. Тогда $A_{x_1^*} = A_{x_2^*} = 0,005$; по теореме сложения абсолютных погрешностей $A_{y^*} = 0,01$ и в соотношении $\sqrt{11} - \sqrt{10} \approx 0,16$ только один знак после запятой верный. Относительная погрешность, как можно посчитать, резко возросла.

Посчитаем искомую величину по другому, используя домножение на сопряженное:

$$\sqrt{11} - \sqrt{10} = \frac{1}{\sqrt{11} + \sqrt{10}} \approx 0,154321$$
$$\sqrt{11} + \sqrt{10} \approx 6,48$$

Оценим погрешность, используя, кроме ранее введенных обозначений, $z = \sqrt{11} + \sqrt{10}$. Тогда $A_{z^*} = 0,01$; $\Delta_{z^*} = \frac{0,01}{6,48} = 0,0016$.

Для оценки погрешности деления $y = 1/z$ используем утверждение

Теорема 2. *При умножении и делении складываются относительные погрешности.*

Доказательство этого утверждения будет дано позже.

Тогда, так как 1 в числителе для y дана точно, $\Delta_{y^*} = \Delta_{z^*} = 0,0016$. Из относительной погрешности можно получить абсолютную

$$A_{y^*} = \Delta_{y^*}|y^*| = 0,0016 \cdot 0,16 = 0,000256.$$

Поэтому в числе y^* три верных знака после запятой.

Таким образом

$$\sqrt{11} - \sqrt{10} \approx 0,154$$

При этом все цифры верные.

2.4. Погрешность функции нескольких переменных

Исследуем погрешность для произвольной функции многих переменных

$$y = f(x_1, x_2, \ldots, x_n)$$

Здесь и в дальнейшем будем предполагать, что функция обладает свойствами, которые обеспечивают дальнейшие действия, например, имеет необходимую гладкость.

Предположим, что заданы приближенные значения аргументов функции $x_1^*, x_2^*, \ldots, x_n^*$ с абсолютными погрешностями $A_{x_1^*}, A_{x_2^*}, \ldots, A_{x_n^*}$. Тогда

$$y^* = f(x_1^*, x_2^*, \ldots, x_n^*).$$

Оценим абсолютную погрешность функции

$$|y - y^*| = |f(x_1, x_2, \ldots, x_n) - f(x_1^*, x_2^*, \ldots, x_n^*)|$$

Сведем разность значений функции векторного аргумента к разности значений функции скалярного аргумента. Введем векторные обозначения

$$X = (x_1, \ldots, x_n) \in R^n,\ X^* = (x_1^*, \ldots, x_n^*) \in R^n$$

Введем также вспомогательную векторную функцию скалярного аргумента:

$$X(t) = X + t(X^* - X).$$

Тогда

$$X(0) = X, X(1) = X^*,$$

$$|f(x_1, x_2, \ldots, x_n) - f(x_1^*, x_2^*, \ldots, x_n^*)| = |f(X(0)) - f(X(1))| =$$

$$= |\phi(0) - \phi(1)|$$

Здесь $\phi(t) = f(X(t))$ – сложная функция. Применим к функции $\phi(t)$ формулу конечных приращений Лагранжа

$$|\phi(0) - \phi(1)| = |\phi'(c)| \cdot |0 - 1|, \ c \in [0; 1]$$

Вычисляя производную функции $\phi(t)$, и производя оценку, получаем

$$|\phi(0) - \phi(1)| = \sum_{i=1}^{n} |\frac{\partial f}{\partial x_i}(X(c))||x_i - x_i^*| \leqslant \sum_{i=1}^{n} |\frac{\partial f}{\partial x_i}(X(c))| A_{x_i^*}$$

Таким образом

$$A_{y^*} = \sum_{i=1}^{n} |\frac{\partial f}{\partial x_i}(X(c))| A_{x_i^*}$$

При малых погрешностях аргументов абсолютная погрешность функции приближенно описывается формулой

$$A_{y^*} \approx \sum_{i=1}^{n} |\frac{\partial f}{\partial x_i}(X^*)| A_{x_i^*}$$

(обычно пишут знак "=" вместо "≈".

Рассмотрим в качестве примера вывод теоремы о погрешности умножения. Пусть

$$y = x_1 x_2$$

Вычисляя частные производные этой функции двух переменных и применяя формулу для погрешностей, получаем

$$A_{y^*} = |x_2^*|A_{x_1^*} + |x_1^*|A_{x_2^*}$$

$$\Delta_{y^*} = \frac{A_{x_1^*}}{|x_1^*|} + \frac{A_{x_2^*}}{|x_2^*|} = \Delta_{x_1^*} + \Delta_{x_2^*}.$$

Аналогично выводится теорема о погрешности деления.

2.5. Погрешность представления вещественных чисел в вычислительном устройстве

Информация о вещественных числах в вычислительном устройстве хранится в виде знака, порядка и мантиссы. Ограниченность разрядов, отводимых для хранения мантиссы приводит к погрешности представления вещественного числа. Формы представления в разных программных средствах различные, но принцип одинаковый, поэтому проиллюстрируем на одной из распространенных, когда на хранение вещественного числа отводится 48 ячеек (бит), из которых одна хранит информацию о знаке числа, восемь ячеек хранят информацию о порядке числа, остальные 39 – о мантиссе числа. Такая форма принята, например, в стандартном представлении вещественного числа в алгоритмическом языке ПАСКАЛЬ. Пусть точное число представимо в виде

$$x = (-1)^s 2^{g-127} 0.a_1 a_2 \dots a_{39} a_{40} \dots,$$

где s – двоичная цифра, g – двоичный порядок, a_i – двоичные цифры нормализованной ($a_1 = 1$) мантиссы.

Приближенное же число имеет в силу ограниченности мантиссы форму

$$x^* = (-1)^s 2^{g-127} 0.a_1 a_2 \dots a_{39}$$

Оценим абсолютную погрешность такого представления

$$|x - x^*| = 2^{g-127}0.\ldots a_{40}\ldots \leq 2^{g-127}0.\ldots 111\ldots \leq$$

$$\leq 2^{g-127}2^{-39} = 2^{g-166},$$

таким образом $A_{x^*} = 2^{g-166}$. Недостаток этой оценки состоит в том, что она зависит от порядка числа.

Оценим относительную погрешность $\Delta_{x^*} = \frac{A_{x^*}}{|x^*|}$

В силу нормализованности мантиссы справедлива оценка

$$2^{g-128} \leq |x^*| \leq 2^{g-127},$$

откуда следует

$$2^{-40} \leq \Delta_{x^*} \leq 2^{-39}.$$

Воспользуемся правилом двоичной тысячи: $2^{10} = 1024 \approx 1000 = 10^3$, тогда $2^{-40} \approx 10^{-12}$, $2^{-39} \approx 5 \cdot 10^{-12}$ и получаем оценку

$$10^{-12} \leq \Delta_{x^*} \leq 0.5 \cdot 10^{-11},$$

таким образом одинадцать верных десятичных цифр после запятой.

Погрешность представления вещественных чисел ведет к различным эффектам, которые нужно учитывать при проведении вычислений.

2.6. Эффект сложения чисел с разным порядком

Пусть в вычислительном устройстве складываются два вещественных числа, имеющих представление

$$x^* = 2^p \cdot m_1$$

и

$$y^* = 2^q \cdot m_2$$

Пусть порядки у них разные, скажем $p >> q$, при сложении происходит выравнивание порядков за счет денормализации мантиссы, поэтому, если разница в порядках очень большая, то может оказаться, что

$$x^* + y^* = x^*$$

Этот эффект приводит к парадоксам, которые "опровергают" известные факты классической математики. Упомянем пару примеров.

Пример 1. При счете по определению гармонический ряд

$$\sum_{n=1}^{\infty} \frac{1}{n}$$

сходится.

В самом деле, частичная сумма ряда

$$S_N = \sum_{n=1}^{N} \frac{1}{n}$$

считается по формуле

$$S_N = S_{N-1} + \frac{1}{N},$$

и, так как второе слагаемое стремится к нулю, а первое не стремится, то начиная с некоторого номера N при реализации счета на некотором вычислительном устройстве будет наблюдаться $S_N = S_{N-1}$, что означает стабилизацию частичных сумм, т.е. "сходимость" ряда.

Пример 2. При счете по определению производные всех функций $f(x)$ "равны нулю", если только $x \neq 0$. В самом деле, по определению

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

Можно организовать цикл для счета предела, взяв, например, $h = 10^{-1}, 10^{-2}, \ldots$ При уменьшении найдется такое h, что с учетом сложения чисел с разным порядком выполняется $x+h = x$ и числитель в определении предела равен нулю, а знаменатель хоть и мал, но не ноль.

2.7. Ускорение сходимости ряда

В качестве иллюстрации влияния вычислительной погрешности рассмотрим задачу о суммирования ряда. Эта задача содержит много однотипных операций, которые легко реализовать с помощью любых программных средств, но в силу того, что погрешности могут накапливаться, результат может сильно искажаться.

Рассмотрим пример: требуется посчитать сумму ряда

$$S = \sum_{n=1}^{\infty} \frac{1}{n^2 + 0,4}$$

с точностью $\varepsilon = 0,5 \cdot 10^{-6}$

Оценим погрешность метода, который состоит в том, что для счета ряда нужно посчитать его частичную сумму

$$S_N = \sum_{n=1}^{N} \frac{1}{n^2 + 0,4} :$$

$$|S - S_N| = \sum_{n=N+1}^{\infty} \frac{1}{n^2 + 0,4}$$

Для оценки остатка ряда используем интегральный признак и оценим интеграл

$$\sum_{n=N+1}^{\infty} \frac{1}{n^2 + 0,4} \leq \int_N^{\infty} \frac{dx}{x^2 + 0,4} \leq \int_N^{\infty} \frac{dx}{x^2} = \frac{1}{N}$$

Потребуем, что погрешность метода не превосходила половины полной погрешности, т.е.

$$|S - S_N| \leq \frac{1}{N} \leq \frac{\varepsilon}{2} = 0,25 \cdot 10^{-6},$$

для этого нужно, чтобы было вычислено $N \geq 4 \cdot 10^6$ слагаемых. Так как при сложении абсолютные погрешность складываются, на подсчет каждого слагаемого (потребуем, чтобы вычислительная погрешность всей суммы

также была равна половине полной) отводится

$$\frac{\varepsilon}{2N} = 0,25 \cdot 10^{-12},$$

т.е. 12 верных знаков после запятой, что невозможно сделать например, если вещественное число имеет форму представления, описанную выше.

Рассмотрим метод, который позволяет ускорить сходимость ряда, резко уменьшить число слагаемых, необходимых для обеспечения заданной точности и, тем самым, уменьшить вычислительную погрешность.

Пусть нужно вычислить исходный ряд

$$A = \sum_{n=1}^{\infty} a_n$$

Ряд

$$B = \sum_{n=1}^{\infty} b_n$$

называется *эталонным* для исходного, если

1. Известна сумма ряда B.

2.

$$\lim_{n\to\infty} \frac{a_n}{b_n} = \lambda \neq 0$$

Если известен эталонный ряд, то исходный ряд можно представить в виде

$$A = \sum_{n=1}^{\infty} (a_n - \lambda b_n) + \lambda B,$$

при этом ряд

$$C = \sum_{n=1}^{\infty} (a_n - \lambda b_n)$$

будет иметь более высокую скорость сходимости чем исходный ряд.

В качестве эталонного ряда могут выступать ряды

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6} = 1,6449340668\ldots$$

$$\sum_{n=1}^{\infty} \frac{1}{n^3} = 1,2020569032\ldots$$

$$\sum_{n=1}^{\infty} \frac{1}{n^4} = \frac{\pi^4}{90} = 1,0823232337\ldots$$

Применим этот прием к ряду, который был рассмотрен в качестве примера. Эталонный ряд для него $\sum\limits_{n=1}^{\infty} \frac{1}{n^2}$, $\lambda = 1$.

Новый ряд

$$C = \sum_{n=1}^{\infty} (\frac{1}{n^2+0,4} - \frac{1}{n^2}) = -\sum_{n=1}^{\infty} \frac{0,4}{n^2(n^2+0,4)}$$

Оценим для него погрешность метода

$$\sum_{n=N+1}^{\infty} \frac{0,4}{n^2(n^2+0,4)} \leq 0,4 \int_N^{\infty} \frac{dx}{x^2(x^2+0,4)} \leq 0,4 \int_N^{\infty} \frac{dx}{x^4} = \frac{0,4}{3N^3}$$

Чтобы теперь обеспечить требуемую погрешность метода

$$\frac{0,4}{3N^3} \leq \frac{\varepsilon}{2} = 0,25 \cdot 10^{-6},$$

достаточно взять порядка $N = 100$ слагаемых и каждое слагаемое нужно считать с точностью до 8 знаков после запятой.

Замечание. При необходимости процедура ускорения сходимости ряда может быть применена еще раз, к уже полученному ряду. Так в примере, после применения эталонного ряда $\sum\limits_{n=1}^{\infty} \frac{1}{n^4}$ число слагаемых во вновь полученном ряде сокращается до десятка.

3. ЧИСЛЕННОЕ РЕШЕНИЕ НЕЛИНЕЙНЫХ УРАВНЕНИЙ

3.1. Постановка задачи

Рассмотрим задачу о численном решении нелинейных уравнений

$$f(x) = 0$$

Несмотря на внешнюю простоту, нелинейные уравнения крайне редко решаются аналитически: даже среди алгебраических эффективно аналитически решаются только квадратные; уравнения третьей и четвертой степени хотя и имеют формулы для аналитического решения (формулы Кардано и Феррари), но в силу их громоздкости в настоящее время такие уравнения предпочитают решать численно; уравнения пятой и выше степени хотя и имеют действительный корень (в силу основной теоремы алгебры), но аналитических формул нет; трансцендентные же уравнения вообще сложны для аналитического решения.

Задач решения нелинейных уравнений состоит из двух этапов: отделение корней и уточнения корней. На первом этапе находится такой отрезок $[a, b]$, на котором существует корень (в дальнейшем будем обозначать его ξ) и он единственен. Эта работа, как правило, проводится аналитически. На втором этапе строится последовательность x_n, сходящаяся к корню. По способу построения этой последовательности и различаются численные методы. Опишем простейшие из них.

3.2. Описание численных методов решения нелинейных уравнений

Метод половинного деления.

Пусть на отрезке $[a, b]$ функция $f(x)$ непрерывна и концах отрезка

принимает разные знаки: $f(a)f(b) < 0$, тогда по теореме Больцано-Коши на этом отрезке существует корень ξ. На этом утверждении и основан метод: находим середину отрезка $c = \frac{a+b}{c}$ и сужаем отрезок так, чтобы на его концах функция принимала разные знаки: если $f(a)f(c) < 0$, то в качестве нового значения правого конца отрезка нужно взять $b = c$, иначе $a = c$. Далее деление отрезка повторяется до тех пор, пока длина отрезка не станет меньше наперед заданной точности ε. Алгоритм всегда сходится, но его недостатком является большое число итераций: число итераций не зависит от функции, а только от длины отрезка.

Замечание. Геометрическая интерпретация этого и других рассмотренных ниже методов обычно рассматривается на практике, поэтому рисунки с иллюстрациями методов приводятся в методической разработке по практическим и лабораторным работам дисциплины "Численные методы".

Метод касательных (Ньютона).

Возьмем начальное приближение x_0 и проведем касательную к графику функции $f(x)$ в этой точке. Точку пересечения касательной с осью абсцисс обозначим за x_1. Далее проводим касательную к графику функции $f(x)$ в точке x_1 и точку пересечения касательной с осью абсцисс обозначим за x_2. Далее процесс повторяется. Из уравнения касательной выводится формула

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

В качестве начального приближения x_0 берут один из концов отрезка, рекомендации по выбору начального приближения этого и других методов будут обсуждаться на практических занятиях.

Как будет доказано позже, метод быстро сходится. Но его недостатком является необходимость использовать производную, а эта аналитическая работа не всегда удобна. Поэтому используют модификации метода.

Упрощенный (модифицированный) метод Ньютона.

Возьмем начальное приближение x_0 и проведем касательную к графику функции $f(x)$ в этой точке. Точку пересечения касательной с осью абсцисс обозначим за x_1. Далее проводим через точку с координатами $(x_1, f(x_1))$ проводим прямую, параллельную первой касательной. Точку пересечения этой прямой с осью абсцисс обозначим за x_2. Далее процесс повторяется по параллельным прямым. Формула следующая

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)}$$

Этот метод требует вычисления всего одного значения производной, но проигрывает методу Ньютона в скорости сходимости.

Метод хорд (неподвижных хорд).

Зафиксируем начальное приближение x_0 и возьмем первое приближение x_1. Через точки с координатами $(x_0, f(x_0))$ и $(x_1, f(x_1))$ на графике функции проведем хорду. Точку пересечения хорды с осью абсцисс обозначим за x_2. Далее проводим хорду через точки с координатами $(x_0, f(x_0))$ и $(x_2, f(x_2))$, пересечения этой хорды с осью абсцисс обозначим за x_3 и так далее. Используя уравнение прямой линии, проходящей через две заданные точки, получаем формулу

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

Метод подвижных хорд.

В отличие от ранее рассмотренных методов метод подвижных хорд двухшаговый – по двум меняющимся значениям считается третье, а затем происходит переадресация. Возьмем два приближения x_0 и x_1. Через точки с координатами $(x_0, f(x_0))$ и $(x_1, f(x_1))$ на графике функции проведем хорду. Точку пересечения хорды с осью абсцисс обозначим за x_2. Далее проводим хорду через точки с координатами $(x_1, f(x_1))$ и $(x_2, f(x_2))$, пе-

ресечения этой хорды с осью абсцисс обозначим за x_3, далее через точки с координатами $(x_2, f(x_2))$ и $(x_3, f(x_3))$ и так далее. Используя уравнение прямой линии, проходящей через две заданные точки, получаем формулу

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

Из более сложных методов рассмотрим

Метод секущих парабол (Майера).

Пусть вторая производная функции на отрезке ограничена $|f''(x)| \leq M$ для всех $x \in [a, b]$ и константа M известна. По значению функции в очередном приближении, по первой производной в этой точке и по M построим производную так, чтобы она пересекала ось абсцисс. Так для случая $f(x_n) > 0$ уравнение этой параболы (ветвями вниз) будет

$$y = f(x_n) + f'(x_n)(x - x_n) - \frac{M}{2}(x - x_n)^2$$

Найдем точки пересечения этой параболы с осью абсцисс, их две, договоримся брать правую, чтобы процесс дал монотонно возрастающую последовательность

$$x_{n+1} = x_n + \frac{f'(x_n) + \sqrt{[f'(x_n)]^2 + 2Mf(x_n)}}{M}$$

Докажем, что между x_n и x_{n+1} корней нет, для этого покажем, что график построенной параболы лежит ниже графика функции $f(x)$. Разложим $f(x)$ по формуле Тейлора в окрестности точки x_n:

$$f(x) = f(x_n) + f'(x_n)(x - x_n) + \frac{f''(c)}{2}(x - x_n)^2,$$

где c некоторая точка между x и x_n. Вычитая из этого уравнения уравнение параболы, получаем

$$y - f(x) = \frac{M + f''(c)}{2}(x - x_n)^2 \geq 0.$$

Таким образом метод дает последовательность, монотонно приближающуюся к корню. Как только нужная точность ε будет достигнута, можно добавить ε к очередному приближению, окажемся справа от корня и пойдем к другому корню (формулы в случае $f(x_n) < 0$ несколько меняются). Достоинство метода состоит в том, что он перебирает все корни на отрезке, если их число конечно, т.е. является глобальным.

Замечание. Если x_n близко к корню, то в формуле метода содержится вычитание близких чисел и, поэтому, нужно её преобразовать, домножив на сопряженное.

3.3. Погрешность методов

Если x_n – очередное приближение, ξ – корень уравнения, то требуемая точность ε достигнута, если выполняется условие $|x_n - \xi| \leq \varepsilon$. Однако это условие нельзя использовать для окончания итерационного процесса, т.к. корень неизвестен. Выведем другую оценку погрешности.

Разложим функцию по формуле Тейлора в окрестности приближения x_n и поставим в разложение корень ξ, получим

$$0 = f(\xi) = f(x_n) + f'(c)(\xi - x_n),\ c \in [\xi, x_n],$$

откуда

$$\xi - x_n = -\frac{f(x_n)}{f'(c)}$$

Предположим, $|f'(x)| \geq m > 0$ на отрезке $[a, b]$, тогда

$$|x_n - \xi| \leq \frac{f(x_n)}{m},$$

и, следовательно, оценку

$$\frac{f(x_n)}{m} \leq \varepsilon,$$

можно взять в качестве условия, обеспечивающего заданную точность.

Замечание. Если оценка производной снизу m неизвестна, то используют эвристическое условие

$$|x_n - x_{n+1}| \leq \varepsilon,$$

которое, однако, не всегда обеспечивает заданную точность.

3.4. Метод простой итерации

Чтобы исследовать главный вопрос – сходимость методов, рассмотрим еще один метод, который играет роль некоторого унифицированного, к которому сводятся все остальные.

Перейдем от исходного уравнения

$$f(x) = 0$$

к эквивалентному уравнению в форме

$$x = \varphi(x)$$

Методом простой итерации для нелинейных уравнений назовем алгоритм

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

Теорема 3. *Пусть функция $\varphi(x)$ непрерывно дифференцируема в окрестности корня ξ, причем выполняется $|\varphi(\xi)| < 1$. Тогда существует такая окрестность $U = (\xi - r, \xi + r)$ корня ξ, что взяв начальное приближение x_0 из этой окрестности, мы получим сходящийся к ξ метод простой итерации.*

Доказательство.

В силу непрерывности производной выполняется $\|\varphi(x)\| \leq q < 1$ в некоторой окрестности $U = (\xi - r, \xi + r)$ корня ξ.

Докажем, что если некоторое приближение $x_k \in U$, то $x_{k+1} \in U$. В самом деле, по теореме Лагранжа

$$|x_{k+1} - \xi| = |\varphi(x_k) - \varphi(\xi)| = |\varphi'(c)||x_{k+1} - \xi| \leq |x_{k+1} - \xi| < r.$$

Таким образом, если взять начальное приближение x_0 из этой окрестности, то все последующие приближения также будут в этой окрестности. Поэтому имеем

$$|x_k - \xi| \leq q|x_{k-1} - \xi| \leq \cdots \leq q^k|x_0 - \xi|$$

откуда следует сходимость $x_k \to \xi$ при $k \to \infty$.

Замечание. Можно доказать также, что если функция $\varphi(x)$ непрерывно дифференцируема в окрестности корня ξ, причем выполняется $|\varphi(\xi)| > 1$, тогда существует такая окрестность $U = (\xi - r, \xi + r)$ корня ξ, что взяв начальное приближение x_0 из этой окрестности, мы получим расходящийся метод простой итерации.

3.5. Сходимость метода Ньютона. Квадратичная сходимость метода Ньютона

В качестве примера доказанного утверждения о сходимости метода простой итерации, докажем сходимость метода Ньютона

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

Метод является частным случаем метода простой итерации, в котором

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

Тогда,

$$\varphi'(x) = 1 - \frac{f'(x)f'(x) - f(x)f''(x)}{[f'(x)]^2}$$

Подставляя корень, в котором $f(\xi) = 0$, получаем

$$\varphi'(\xi) = 0.$$

Таким образом, метод Ньютона сходится, если "удачно", т.е. в нужной окрестности корня взять начальное приближение.

Более того, метод Ньютона обладает квадратичной скоростью сходимости, т.е. расстояние от последующего приближения до корня убывает пропорционально квадрату расстояния от последующего приближения до корня.

В самом деле,

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

Но используя разложение по формуле Тейлора, получаем

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

тогда

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

Если $|f'(x)| \geq m > 0$, $|f''(x)| \leq M$ на отрезке $[a, b]$, то

$$|\xi - x_{n+1}| \leq \frac{1}{2}\frac{M|\xi - x_n|^2}{m}$$

Введем величину

$$r_n = \frac{1}{2}\frac{M|\xi - x_n|}{m},$$

тогда

$$r_{n+1} = \frac{1}{2}\frac{M|\xi - x_{n+1}|}{m} \leq \frac{M}{2m}\frac{1}{2}\frac{M|\xi - x_n|^2}{m} = r_n^2,$$

$$r_{n+1} \leq r_n^2 \leq r_{n-1}^4 \leq \cdots \leq r_0^{2^{n+1}},$$

что и означает квадратичную скорость сходимости.

4. ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ ЗАДАЧ ЛИНЕЙНОЙ АЛГЕБРЫ

Среди многочисленных задач линейной алгебры наибольшее значение имеет задача о решении линейных систем, к которой сводятся другие задачи. Рассмотрим эту задачу.

4.1. Численные методы решения линейных систем. Классификация методов

Пусть дана линейная неоднородная система

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n. \end{cases}$$

Введем обозначения

$$A = \begin{pmatrix} a_{11} & \ldots & a_{1n} \\ . & \ldots & . \\ a_{n1} & \ldots & a_{nn} \end{pmatrix}, \; x = \begin{pmatrix} x_1 \\ . \\ x_n \end{pmatrix}, \; b = \begin{pmatrix} b_1 \\ . \\ b_n \end{pmatrix},$$

тогда система запишется в виде

$$Ax = b$$

Будем предполагать, что определитель матрицы A отличен от нуля, т.е. система имеет единственное решение.

Все численные методы делятся на две группы: точные и итерационные.

В *точных методах* за конечное число шагов достигается точное решение системы. В *итерационных методах* за конечное число шагов достигается лишь некоторое приближение к точному решению.

Рассмотрим сначала группу точных методов. Основной критерий при выборе точных методов – число операций, необходимых для решения системы. При этом под количеством операций обычно понимают количество операций умножения, т.к. на современных вычислительных устройствах операции сложения и вычитания выполняются настолько быстро по сравнению с операцией умножения, что ими пренебрегают в расчетах затрат времени, а операция деления встречается крайне редко, её стараются избегать.

Известный из курса алгебры метод Крамера для решения линейных систем, относящийся к точным методам, в этом смысле крайне нерационален, число операций растет как факториал относительно размерности n системы. Даже на современных компьютерах для решения систем размерности 20 потребуется 10^8 лет. А приходится решать системы и гораздо большей размерности.

Совсем другая ситуация с методом исключения неизвестных Гаусса, который также относится к точным методам. Там число операций имеет порядок n^3 и для решения той же задачи потребуется менее секунды. Классическая схема исключения неизвестных обладает недостатком: многочисленные переходы от одной матрицы к другой. Она предназначена для аналитического счета, но не для программирования алгоритма. Рассмотрим вариант этого метода, который позволяет осуществить переход от исходной матрицы к последней за один шаг.

4.2. Компактная схема Гаусса

Для того, чтобы изложение алгоритма было понятнее, распишем его для системы размерности четыре. Пусть решается система

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 = a_{15}, \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 = a_{25}, \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + a_{34}x_4 = a_{35}, \\ a_{41}x_1 + a_{42}x_2 + a_{43}x_3 + a_{44}x_4 = a_{45}. \end{cases}$$

Таким образом

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix}$$

Рассмотрим вспомогательную задачу факторизации матрицы A, т.е. представление матрицы в виде

$$A = BC,$$

где матрица B – левая треугольная, C – правая треугольная, по главной диагонали которой стоят единицы:

$$B = \begin{pmatrix} b_{11} & 0 & 0 & 0 \\ b_{21} & b_{22} & 0 & 0 \\ b_{31} & b_{32} & b_{33} & 0 \\ b_{41} & b_{42} & b_{43} & b_{44} \end{pmatrix}, \; C = \begin{pmatrix} 1 & c_{12} & c_{13} & c_{14} \\ 0 & 1 & c_{23} & c_{24} \\ 0 & 0 & 1 & c_{34} \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Запись этих двух матриц можно объединить в одну (подрузамевает-

ся, но не пишется, что на главной диагонали матрицы C стоят единицы)

$$A^* = \begin{pmatrix} b_{11} & c_{12} & c_{13} & c_{14} \\ b_{21} & b_{22} & c_{23} & c_{24} \\ b_{31} & b_{32} & b_{33} & c_{34} \\ b_{41} & b_{42} & b_{43} & b_{44} \end{pmatrix}$$

Расписывая элементы первого столбца произведения матриц B и C, последовательно получаем

$$a_{11} = b_{11}c_{11} = b_{11},\ a_{21} = b_{21},\ a_{31} = b_{31},\ a_{41} = b_{41},$$

т.е. первый столбец "новой" матрицы A^* совпадает с первым столбцом "старой" матрицы A.

Расписывая элементы первой строки произведения матриц B и C, получаем

$$a_{12} = b_{11}c_{12},\ a_{13} = b_{11}c_{13},\ a_{14} = b_{11}c_{14},$$

откуда

$$c_{12} = \frac{a_{12}}{b_{11}},\ c_{13} = \frac{a_{13}}{b_{11}},\ c_{14} = \frac{a_{14}}{b_{11}},$$

т.е. элементы первой строки "новой" матрицы A^* получаются из соответствующих элементов "старой" матрицы A делением на диагональный элемент "новой" матрицы.

Расписывая элементы второго столбца произведения матриц B и C, получаем

$$a_{22} = b_{21}c_{12} + b_{22},\ a_{32} = b_{31}c_{12} + b_{32},\ a_{42} = b_{41}c_{12} + b_{42},$$

откуда

$$b_{22} = a_{22} - b_{21}c_{12},\ b_{32} = a_{32} - b_{31}c_{12},\ b_{42} = a_{42} - b_{41}c_{12},$$

т.е. элементы второго столбца матрицы B получаются из соответствующих элементов "старой" матрицы A, из которых нужно вычесть произведение уже подсчитанных соответствующих (стоящих в той же строке и колонке) элементов "новой" матрицы.

Расписывая элементы второй строки произведения матриц B и C, получаем

$$a_{23} = b_{21}c_{13} + b_{22}c_{23},\ a_{24} = b_{21}c_{14} + b_{22}c_{24},$$

откуда

$$c_{23} = \frac{(a_{23} - b_{21}c_{13})}{b_{22}},\ c_{24} = \frac{(a_{24} - b_{21}c_{14})}{b_{22}},$$

т.е. элементы второй строки матрицы C получаются из соответствующих элементов "старой" матрицы A, из которых нужно вычесть произведение уже подсчитанных соответствующих (стоящих в той же строке и колонке) элементов "новой" матрицы и затем поделить на диагональный элемент "новой" матрицы.

Аналогичным образом последовательно получаются оставшиеся элементы новой матрицы:

$$b_{33} = a_{33} - b_{31}c_{13} - b_{32}c_{23},\ b_{43} = a_{43} - b_{41}c_{13} - b_{42}c_{23},$$

$$c_{34} = \frac{(a_{34} - b_{31}c_{14} - b_{32}c_{24})}{b_{33}},$$

$$b_{44} = a_{44} - b_{41}c_{14} - b_{42}c_{24} - b_{43}c_{34}$$

Таким образом, можем сформулировать общие правила.

Последовательно считаются столбец, затем строка, снова столбец, снова строка и т.д. "новой" матрицы. Элементы матрицы B, т.е.элементы "новой" матрицы, стоящие ниже и на главной диагонали, получаются из соответствующих элементов "старой" матрицы A, из которых нужно вычесть произведение уже подсчитанных соответствующих (стоящих в той же строке и колонке) элементов "новой" матрицы. Элементы матрицы C,

т.е.элементы "новой" матрицы, стоящие выше главной диагонали, получаются из соответствующих элементов "старой" матрицы A, из которых нужно вычесть произведение уже подсчитанных соответствующих (стоящих в той же строке и колонке) элементов "новой" матрицы и всё поделить на диагональный элемент новой матрицы.

Общие формулы следующие

$$b_{ij} = a_{ij} - \sum_{k=1}^{j-1} b_{ik}c_{kj},\ j = 1, \ldots, n,\ i = j, \ldots, n,$$

$$c_{ij} = \frac{a_{ij} - \sum\limits_{k=1}^{i-1} b_{ik}c_{kj}}{b_{ii}}, i = 1, \ldots, n-1,\ j = i+1, \ldots, n$$

Вернемся к решению линейной системы $Ax = b$. Предположим, что решена задача факторизации $A = BC$, тогда получаем $BCx = b$ и, обозначая $Cx = y$, сводим исходную задачу к задаче решения двух систем с треугольными матрицами

$$\begin{cases} By = b \\ Cx = y \end{cases}$$

Распишем для нашего случая $n = 4$ первую из этих систем

$$\begin{cases} b_{11}y_1 = a_{15} \\ b_{21}y_1 + b_{22}y_2 = a_{25} \\ b_{31}y_1 + b_{32}y_2 + b_{33}y_3 = a_{35} \\ b_{41}y_1 + b_{42}y_2 + b_{43}y_3 + b_{44}y_4 = a_{45} \end{cases}$$

Отсюда

$$\begin{cases} y_1 = \frac{a_{15}}{b_{11}} \\ y_2 = \frac{a_{25} - b_{21} y_1}{b_{22}} \\ y_3 = \frac{(a_{35} - b_{31} y_1 - b_{32} y_2)}{b_{33}} \\ y_4 = \frac{(a_{45} - b_{41} y_1 - b_{42} y_2 - b_{43} y_3)}{b_{44}} \end{cases}$$

т.е. формулы для y_i такие же, что и для c_{i5}. Тогда введя обозначения $c_{i5} = y_i$, мы можем унифицировать действия с основной матрицей и столбцом правых частей исходной системы. Введем "старую" расширенную матрицу $\bar{A}$ и "новую" расширенную матрицу $\bar{A}^*$.

$$\bar{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \end{pmatrix}$$

$$\bar{A}^* = \begin{pmatrix} b_{11} & c_{12} & c_{13} & c_{14} & c_{15} \\ b_{21} & b_{22} & c_{23} & c_{24} & c_{25} \\ b_{31} & b_{32} & b_{33} & c_{34} & c_{35} \\ b_{41} & b_{42} & b_{43} & b_{44} & b_{45} \end{pmatrix}$$

Переход от элементов "старой" матрицы к элементам "новой" матрицы осуществляется по указанным выше правилам. Это прямой ход в схеме исключения неизвестных, представляющий собой тройной цикл с числом операций, пропорциональным n^3.

Обратный ход алгоритма состоит в решении системы

$$Cx = y$$

и представляет собой двойной цикл.

Замечание 1. Как показывают формулы алгоритма, компактная схема Гаусса осуществима тогда и только тогда, когда все диагональные элементы b_{ii} отличны от нуля. В терминах исходной матрицы это выполняется тогда и только тогда, когда все главные миноры матрицы A отличны от нуля.

Замечание 2. Если элементы b_{ii} (главные элементы) отличны от нуля, но близки к нулю по абсолютной величине, то при делении на малую величину может возникнуть большая погрешность. Чтобы избежать этого явления, применяют модификации компактной схемы Гаусса: схему с выбором главного элемента по колонке, схему с выбором главного элемента по строке, схему с выбором главного элемента по всей матрице.

Замечание 3. С помощью компактной схемы можно считать определители квадратных матриц. В силу того, что определитель произведения матриц равен произведению определителей матриц, а матрицы B и C треугольные, получаем

$$det(A) = det(BC) = det(B)det(C) = det(B) = \prod_{i=1}^{n} b_{ii}$$

Компактная схема Гаусса является одним из самых эффективных алгоритмов вычисления определителей и этот алгоритм содержится во многих пакетах прикладных программ.

Замечание 4. Компактной схемой Гаусса можно эффективно решать несколько систем с одной и той же основной матрицей A и с разными правыми частями. В этом случае в расширенную матрицу $\bar{A}$ дописываются новые столбцы правых частей, а матрица B будет одна и та же, т.е. число операций возрастет не сильно.

Замечание 5. С помощью компактной схемы можно считать обратные матрицы. В самом деле, пусть $X = A^{-1}$, тогда матрица X является

решением матричного уравнения

$$AX = E,$$

E – единичная матрица. Обозначим столбцы матрицы X через x^i, а через e^i – столбец, на i месте у которого стоит 1, а остальные координаты равны нулю. Тогда получаем n систем линейных уравнений

$$Ax^i = e^i,$$

с одной и той же основной матрицей. Эти системы эффективно, согласно замечанию 4, одновременно решаются компактной схемой Гаусса.

4.3. Трехдиагональная прогонка

Кроме компактной схемы Гаусса, существует много других точных методов. Некоторые из них используют специфику системы. Так метод квадратного корня, который использует симметричность матрицы A, в два раза сокращает число операций. Но особо эффективным и потому часто применяемым является алгоритм прогонки, который использует трехдиагональность матрицы A. Матрица называется *трехдиагональной*, если у неё отличны от нуля только элементы, стоящие на главной диагонали и на двух соседних.

Рассмотрим систему с трехдиагональной матрицей

$$\begin{cases} b_1x_1 + c_1x_2 = d_1 \\ a_2x_1 + b_2x_2 + c_2x_3 = d_2 \\ \cdots\cdots\cdots \\ a_ix_{i-1} + b_ix_i + c_ix_{i+1} = d_i \\ \cdots\cdots\cdots \\ a_nx_{n-1} + b_nx_n = d_n \end{cases}$$

Если выразить из первого уравнения x_1 через x_2 и подставить во второе, то x_2 выразится линейно через x_3 и т.д. Введем линейную связь предыдущего неизвестного через последующее (обратный ход прогонки)

$$x_{i-1} = \lambda_i x_i + \mu_i,$$

λ_i и μ_i называются *прогоночными коэффициентами.* Чтобы их определить, подставим это выражение в i-ое уравнение

$$a_i(\lambda_i x_i + \mu_i) + b_i x_i + c_i x_{i+1} = d_i,$$

откуда

$$x_i = -\frac{c_i}{a_i\lambda_i + b_i} x_{i+1} + \frac{d_i - a_i\mu_i}{a_i\lambda_i + b_i}$$

Получили связь между предыдущим и последующим неизвестным и, согласно обозначениям,

$$\lambda_{i+1} = -\frac{c_i}{a_i\lambda_i + b_i},\ \mu_{i+1} = \frac{d_i - a_i\mu_i}{a_i\lambda_i + b_i}$$

Для того, чтобы начинать счет прогоночных коэффициентов по этим формулам, выразим x_1 из первого уравнения системы:

$$x_1 = -\frac{c_1}{b_1} x_2 + \frac{d_1}{b_1},$$

откуда

$$\lambda_2 = -\frac{c_1}{b_1},\ \mu_2 = \frac{d_1}{b_1}$$

Сравнивая эти формулы с общими формулами для прогоночных коэффициентов, получаем, что цикл нужно начинать с

$$\lambda_1 = 0,\ \mu_1 = 0.$$

Для того, чтобы вывести формулу, с которой нужно начинать обратный ход прогонки, подставим x_{n-1} в последнее уравнение системы:

$$a_n(\lambda_n x_n + \mu_n) + b_n x_n = d_n,$$

откуда

$$x_n = \frac{d_n - a_n\mu_n}{a_n\lambda_n + b_n} = \mu_{n+1},$$

т.е. обратный ход нужно начинать с формулы

$$x_n = \mu_{n+1}.$$

Алгоритм прогонки очень эффективен, т.к. число операций линейно относительно размерности системы n, этот факт дает возможность решать системы очень большой размерности.

Установим условия, при которых прогонка осуществима.

Будем говорить, что матрица обладает *диагональным преобладанием*, если для каждой строки модуль диагонального элемента больше суммы модулей всех других элементов этой строки.

В нашем случае диагональное преобладание матрицы A означает, что

$$\Delta_i = |b_i| - |a_i| - |c_i| > 0$$

для всех $i = 1, 2, \ldots, n$.

Лемма 1. Пусть для основной матрицы системы выполнены условия диагонального преобладания. Тогда выполняется

$$|\lambda_i| < 1$$

для всех $i = 1, 2, \ldots, n$.

Доказательство. Проверим утверждение леммы индукцией по i.

База индукции выполняется, так как $\lambda_1 = 0$.

Предположим, что $|\lambda_i| < 1$ докажем, что $|\lambda_{i+1}| < 1$. В силу диагонального преобладания имеем оценку

$$|a_i\lambda_i + b_i| \geq |b_i| - |a_i||\lambda_i| \geq |b_i| - |a_i| = \Delta_i + |c_i|$$

По формулам прогоночных коэффициентов

$$|\lambda_{i+1}| = \frac{|c_i|}{|a_i\lambda_i + b_i|} \leq \frac{|c_i|}{\Delta_i + |c_i||} < 1$$

Доказательство. В условиях диагонального преобладания прогонка осуществима. Это следует из того, что в процессе доказательства леммы мы показали, что модуль знаменателя в формулах для прогоночных коэффициентов оценивается снизу положительной величиной.

Докажем также важное утверждение об оценке решений системы с трехдиагональной матрицей. Это утверждение часто используется для доказательства устойчивости алгоритмов, в которых используется прогонка.

Лемма 2. Пусть для основной матрицы системы выполнены условия диагонального преобладания. Тогда выполняется оценка координат решения системы

$$\max_{1\leq i\leq n} |x_i| \leq \max_{1\leq i\leq n} \frac{|d_i|}{\Delta_i}$$

Доказательство. Пусть j – тот номер, на котором достигается

$$\max_{1\leq i\leq n} |x_i| = |x_j|.$$

Рассмотрим j-ое уравнение системы

$$a_j x_{j-1} + b_j x_j + c_j x_{j+1} = d_j,$$

откуда

$$b_j x_j = d_j - a_j x_{j-1} - c_j x_{j+1},$$

$$|b_j||x_j| \leq |d_j| + |a_j||x_{j-1}| + |c_j||x_{j+1}| \leq |d_j| + (|a_j| + |c_j|)|x_j|.$$

Перенося последнее слагаемое в левую часть неравенства, получаем

$$\Delta_j|x_j| \leq |d_j|,$$

или

$$\max_{1\leq i\leq n} |x_i| = |x_j| \leq \frac{|d_j|}{\Delta_j} \leq \max_{1\leq i\leq n} \frac{|d_i|}{\Delta_i},$$

что и требовалось доказать.

4.4. Нормы матриц

Прежде чем рассматривать итерационные методы решения линейных систем, рассмотрим вспомогательный аппарат: понятие нормы квадратной матрицы. Это понятие вводится по аналогии к норме вектора. Вспомним это определение.

Пусть x – n-мерный вектор с координатами x_i. Нормой вектора называется поставленное в соответствие x число $\|x\|$, удовлетворяющее свойствам:

1. Для всякого вектора x выполняется $\|x\| \geq 0$, причем $\|x\| = 0$ тогда и только тогда, когда вектор x нулевой.

2. Для всякого вектора x и скаляра λ выполняется $\|\lambda x\| = |\lambda|\|x\|$.

3. Для всяких векторов x и y выполняется $\|x + y\| \leq \|x\| + \|y\|$.

Среди различных видов норм наиболее распространенными являются

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

$$\|x\|_2 = (\sum_{i=1}^{n} x_i^2)^{1/2}$$

$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$$

Пусть A – квадратная матрица размерности $n \times n$ с элементами a_{ij}. *Нормой матрицы* называется поставленное в соответствии A число $\|A\|$, удовлетворяющее свойствам:

1. Для всякой матрицы A выполняется $\|A\| \geq 0$, причем $\|A\| = 0$ тогда и только тогда, когда матрица A нулевая.

2. Для всякой матрицы A и скаляра λ выполняется $\|\lambda A\| = |\lambda|\|A\|$.

3. Для всяких матриц A и B выполняется $\|A + B\| \leq \|A\| + \|B\|$.

Это общее определение нормы матрицы. С другой стороны, при выбранном базисе в векторном пространстве матрице соответствует линейное отображение $x \to Ax$, поэтому на матрицу можно смотреть как на линейный оператор и ввести его норму:

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

Это определение зависит от способа введения нормы вектора x и называется *подчиненной нормой матрицы.*

Не всякая норма матрицы является подчиненной.

Кроме указанных трех свойств, подчиненная норма матрицы удовлетворяет еще следующим свойствам:

4. Для всякой матрицы A и вектора x выполняется $\|Ax\| \leq \|A\|\|x\|$.

5. Для всяких матриц A и B выполняется $\|AB\| \leq \|A\|\|B\|$.

Укажем без доказательства формулы, которыми описываются нормы матриц, подчиненные наиболее распространенным векторным нормам.

$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{1 \leq i \leq n} |a_{ij}|$$

(поколонная норма);

$$\|A\|_\infty = \max_{1 \leq i \leq n} \sum_{1 \leq j \leq n} |a_{ij}|$$

(построчная норма);

$$\|A\|_2 = \sqrt{\max_i \lambda_i(A^T A)}$$

(евклидова норма).

В определении евклидовой нормы A^T – означает транспонированную матрицу, $\lambda_i(B)$ – собственные числа матрицы. Определение евклидовой нормы корректно, т.к. $A^T A$ – симметричная, неотрицательно определённая матрица, поэтому её собственные числа вещественны и неотрицательны и под корнем стоит неотрицательное число.

В отличие от поколонной и построчной норм, евклидова норма определяется не очень конструктивно из-за операции вычисления собственных чисел, поэтому часто используют оценку

$$\|A\|_2 \leq \sqrt{\sum_{i,j}(a_{ij})^{1/2}}$$

Кроме того, если матрица A – симметричная, то

$$\|A\|_2 = \max_i |\lambda_i(A)|.$$

В дальнейшем будем рассматривать только подчиненные нормы.

4.5. Метод простой итерации для решения систем линейных уравнений. Критерий сходимости

Преобразуем исходную систему

$$Ax = b$$

к виду

$$x = Bx + C$$

Методом простой итерации назовем алгоритм

$$x^{(k+1)} = Bx^{(k)} + C,$$

где $x^{(k)}$ – k-ое приближение вектора x к решению.

Фактически все итерационные методы можно записать в таком виде и главный вопрос в исследовании метода – условия его сходимости. Распишем несколько итераций, взяв начальное приближение $x^{(0)}$. Тогда

$$x^{(1)} = Bx^{(0)} + C,$$

$$x^{(2)} = Bx^{(1)} + C = B^2x^{(0)} + (E + B)C,$$

$$x^{(m)} = B^mx^{(0)} + (E + B + B^2 + \cdots + B^{m-1})C.$$

Вопрос сходимости упирается в условия сходимости матричного ряда в этой формуле.

Справедливо следующее фундаментальное утверждение о сходимости метода простой итерации для линейных систем:

Теорема 4. *Метод простой итерации сходится при любом начальном приближении тогда и только тогда, когда все собственные числа матрицы B удовлетворяют условию*

$$|\lambda_i(B)| < 1$$

Доказательство разобьем на ряд вспомогательных утверждений.

Сначала дадим определение.

ε-жордановой формой матрицы называется квадратная матрица, у которой диагональные клетки являются ε-жордановыми, а остальные клетки нулевые.

ε-жордановой клеткой называется клетка, у которой на главной диагонали стоит собственное число матрицы, ниже по соседней диагонали стоят числа ε, а все остальные элементы нулевые:

$$\begin{pmatrix} \lambda & 0 & \cdots & 0 \\ \varepsilon & \lambda & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda \end{pmatrix}$$

Лемма 1. Всякая квадратная матрица B для всякого $\varepsilon > 0$ может быть приведена к ε-жордановой форме.

Доказательство.

Возьмем квадратную матрицу B и любое $\varepsilon > 0$, обозначим $A = \frac{1}{\varepsilon}B$. По одному из основных утверждений линейной алгебры всякую квадратную матрицу A можно привести к жордановой форме (которая отличается от ε-жордановой формы тем, что в клетках на диагонали ниже главной стоят единицы), т.е. найдется невырожденная матрица T, такая, что матрица $G = T^{-1}AT$ – жорданова. Умножим это соотношение на ε, получим

$$\varepsilon G = T^{-1}(\varepsilon A)T = T^{-1}BT.$$

Из определения собственного числа $Bh = \lambda(B)h$, (h – собственный вектор) вытекает, что введенная матрица A имеет собственные числа $\lambda(A) = \frac{1}{\varepsilon}\lambda(B)$, такие же собственные числа имеет жорданова матрица G. Но тогда у матрицы $G_\varepsilon = \varepsilon G$ такие же собственные числа, что и у матрицы B, т.е. матрица G_ε является ε-жордановой формой матрицы B.

Лемма 2. Для того, чтобы $B^m \to 0$ при $m \to \infty$, необходимо и достаточно, чтобы для всех собственных чисел матрицы B выполнялось условие $|\lambda_i(B)| < 1$.

Доказательство.

Так как $G_\varepsilon^2 = T^{-1}BTT^{-1}BT = T^{-1}B^2T$, то $G_\varepsilon^m = T^{-1}B^mT$, и обратно $B^m = TG_\varepsilon^mT^{-1}$. Из свойств подчиненной нормы вытекает, что

$$\|G_\varepsilon^m\| \le \|T^{-1}\|\|B^m\|\|T\|$$

и

$$\|B^m\| \le \|T\|\|G_\varepsilon^m\|\|T^{-1}\,|$$

матрицы T и T^{-1} – невырожденные, поэтому сходимость к нулевой матрице матрицы B^m эквивалентна сходимости к нулевой матрице матрицы G_ε^m.

Пусть выполняется $|\lambda_i(B)| < 1$ для всех собственных чисел матрицы B.

Тогда найдется $\varepsilon > 0$, такое, что

$$\max_i |\lambda_i(B)| + \varepsilon < 1$$

Рассмотрим построчную норму матрицы G_ε:

$$\|G_\varepsilon\|_\infty = \max_i |\lambda_i(G_\varepsilon)| + \varepsilon = \max_i |\lambda_i(B)| + \varepsilon = q < 1$$

Так как $\|G_\varepsilon^m\| \leq \|G_\varepsilon\|^m$, то это означает сходимость к нулевой матрице матрицы G_ε^m и, следовательно матрицы B^m.

Обратно, пусть $G_\varepsilon^m \to 0$ при $m \to \infty$.

Матрица G_ε^m состоит из клеток, которые имеет следующую структуру: на главной диагонали стоят числа $\lambda_i(B)^m$. Так как она сходится к нулевой, то $\lambda_i(B)^m \to 0$, что означает выполнимость условия $|\lambda_i(B)| < 1$ для всех собственных чисел матрицы B.

Лемма 3. Матричный ряд $E + B + B^2 + \cdots + B^m + \cdots$ сходится тогда и только тогда, когда для всех собственных чисел матрицы B выполняется условие $|\lambda_i(B)| < 1$. При этом

$$\sum_{m=0}^{\infty} B^m = (E - B)^{-1}$$

Доказательство.

Пусть матричный ряд сходится, тогда общий член ряда стремится к нулю, т.е. $B^m \to 0$ при $m \to \infty$, а тогда по лемме 2 выполняется $|\lambda_i(B)| < 1$ для всех собственных чисел матрицы B.

Обратно, пусть выполняется $|\lambda_i(B)| < 1$ для всех собственных чисел матрицы B.

Рассмотрим тождество

$$(E + B + B^2 + \cdots + B^{m-1})(E - B) = E - B^m,$$

которое проверяется раскрытием скобок.

Покажем, что матрица $E - B$ обратима. В самом деле, если бы это было не так, то система линейных однородных уравнений

$$(E - B)x = 0$$

имела бы нетривиальное решение x, а это означало бы, что вектор x собственный для матрицы B с собственным числом $\lambda(B) = 1$, что противоречит условию $|\lambda_i(B)| < 1$. Итак матрица $E - B$ обратима и указанное тождество можно умножить справа на $(E - B)^{-1}$.

В полученном соотношении

$$E + B + B^2 + \cdots + B^{m-1} = (E - B^m)(E - B)^{-1}$$

перейдем к пределу при $m \to \infty$. По лемме 2 выполняется $B^m \to 0$ и правая часть соотношения стремится к пределу $(E - B)^{-1}$. Тогда сумма матричного ряда в левой части соотношения тоже имеет предел и ряд сходится к этому пределу:

$$\sum_{m=0}^{\infty} B^m = (E - B)^{-1}$$

Доказательство теоремы.

Вернемся к соотношению

$$x^{(m)} = B^m x^{(0)} + (E + B + B^2 + \cdots + B^{m-1})C.$$

Пусть выполняется $|\lambda_i(B)| < 1$ для всех собственных чисел матрицы B, тогда по лемме 2 $B^m \to 0$ и первое слагаемое в правой части этого соотношения стремится к нулю, а по лемме 3 частичная сумма матричного ряда стремится к $(E - B)^{-1}$. Тогда правая и левая части соотношения имеют предел

$$x = (E - B)^{-1}C.$$

Умножая на $(E - B)$ и преобразуя, получаем, что вектор x удовлетворяет соотношению

$$x = Bx + C$$

Предположим теперь, что метод простой итерации сходится при любом начальном приближении, в частности при $x^{(0)}$. Тогда последовательность

$$x^{(m)} = (E + B + B^2 + \cdots + B^{m-1})C$$

сходится, поэтому матричный ряд сходится, и по лемме 3 выполняется $|\lambda_i(B)| < 1$ для всех собственных чисел матрицы B.

4.6. Достаточные условия метода простой итерации для решения систем линейных уравнений

Приведенные необходимые и достаточные условия сходимости метода простой итерации не совсем удобны: они связаны со сложной задачей нахождения всех собственных значений матрицы. Следующее вспомогательное утверждение дает путь получения многих простых достаточных условий сходимости.

Лемма. Модуль собственного числа матрицы не превосходит нормы матрицы:

$$|\lambda(B)| \leq \|B\|$$

Доказательство.

Пусть λ – собственное число матрицы B, а x – соответствующий собственный вектор, тогда по определению $Bx = \lambda x$. Из свойств подчиненной нормы вытекает

$$\|B\|\|x\| \geq \|Bx\| = |\lambda x| = |\lambda|\|x\|.$$

Так как собственный вектор x ненулевой, его норма не ноль, поэтому можно в неравенстве сократить на $\|x\|$, получим требуемое.

Как следствие из этого утверждения и критерия сходимости метода простой итерации для линейных систем вытекает

Теорема 5. *Если* $\|B\| < 1$, *то метод простой итерации сходится при любом начальном приближении.*

Доказательство получается из неравенства

$$|\lambda(B)| \leq \|B\| < 1$$

В частности, метод простой итерации сходится, если выполняется одно из следующих условий:

1). Для любого $j = 1, \ldots, n$

$$\sum_{i=1}^{n} |b_{ij}| < 1$$

2).

$$\sum_{i,j=1}^{n} b_{ij}^2 < 1$$

3). Для любого $i = 1, \ldots, n$

$$\sum_{j=1}^{n} |b_{ij}| < 1$$

4.7. Метод Якоби

Конкретные итерационные методы различаются по способу перехода от исходной линейной системы к виду, удобному для организации итерационного процесса, сходимость которого определяется матрицей B. Рассмотрим простейший способ такого перехода.

Пусть дана система

$$Ax = b$$

или в координатной форме

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n. \end{cases}$$

Выразим x_1 из первого уравнения, x_2 из второго уравнения и т.д., x_n из последнего уравнения (предполагается, что диагональные элементы матрицы A отличны от нуля), получим систему вида

$$\begin{cases} x_1 = -\frac{a_{12}}{a_{11}}x_2 - \cdots - \frac{a_{1n}}{a_{11}}x_n + \frac{b_1}{a_{11}}, \\ x_2 = -\frac{a_{21}}{a_{22}}x_1 - \cdots - \frac{a_{2n}}{a_{22}}x_n + \frac{b_2}{a_{22}}, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x_n = -\frac{a_{n1}}{a_{nn}}x_1 - \frac{a_{n2}}{a_{nn}}x_2 - \cdots + \frac{b_n}{a_{nn}} \end{cases}$$

или в векторной форме

$$x = Bx + C,$$

где матрица B имеет вид

$$B = \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} & \ldots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & \ldots & -\frac{a_{2n}}{a_{22}} \\ . & . & \ldots & . \\ -\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & \ldots & 0 \end{pmatrix}$$

Методом Якоби назовем итерационный процесс

$$\begin{cases} x_1^{(k+1)} = -\frac{a_{12}}{a_{11}}x_2^{(k)} - \cdots - \frac{a_{1n}}{a_{11}}x_n^{(k)} + \frac{b_1}{a_{11}}, \\ x_2^{(k+1)} = -\frac{a_{21}}{a_{22}}x_1^{(k)} - \cdots - \frac{a_{2n}}{a_{22}}x_n^{(k)} + \frac{b_2}{a_{22}}, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x_n^{(k+1)} = -\frac{a_{n1}}{a_{nn}}x_1^{(k)} - \frac{a_{n2}}{a_{nn}}x_2^{(k)} - \cdots + \frac{b_n}{a_{nn}} \end{cases}$$

Для того, чтобы изучить условия его сходимости, представим исходную матрицу A в виде $A = L+D+R$, где L – содержит элементы исходной матрицы ниже главной диагонали, все остальные элементы нулевые, D – диагональная матрица, R – содержит элементы исходной матрицы выше главной диагонали. Тогда сделанные преобразования запишутся в матричном виде следующим образом:

$$(L + D + R)x = b,\ Dx = -(L + R)x + b,\ x = -D^{-1}(L + R)x + D^{-1}b,$$

и матрица B в методе простой итерации определяется соотношением

$$B = -D^{-1}(L + R)$$

Теорема 6. *Пусть $a_{ii} \neq 0$ для всех $i = 1, \ldots, n$. Тогда для сходимости метода Якоби необходимо и достаточно, чтобы все корни уравнения*

$$det(L + \lambda D + R) = 0$$

удовлетворяли соотношению $|\lambda| < 1$.

Доказательство.

Согласно доказанному критерию сходимости метода простой итерации, метод сходится тогда и только тогда, когда все собственные числа матрицы B по модулю меньше единицы. Выпишем характеристическое уравнение, учитывая конкретный вид матрицы в методе Якоби и проделаем

преобразования

$$0 = det(\lambda E - B) = det(\lambda E + D^{-1}(L + R)) =$$

$$= det(D^{-1})det(L + \lambda D + R),$$

откуда следует заключение теоремы.

Теорема 7. *Если исходная матрица A имеет диагональное преобладание, то метод Якоби сходится.*

Доказательство.

В условиях диагонального преобладания матрицы A выпишем построчную норму матрицы B:

$$\|B\|_\infty = \max_{1\le i\le n} \sum_{1\le j\le n} |b_{ij}| = \max_{1\le i\le n} \sum_{1\le j\le n, j\ne i} \frac{|a_{ij}|}{|a_{ii}|} < 1,$$

поэтому, в силу достаточного условия, метод сходится.

4.8. Метод Гаусса-Зейделя

Посмотрим на покоординатные формулы метода Якоби. Когда считается вторая координата $k+1$-го приближения, уже известно $x_1^{(k+1)}$, поэтому в формулах естественно использовать уже посчитанные приближения координат. Получаем метод, называемый *методом Гаусса-Зейделя*

$$\begin{cases} x_1^{(k+1)} = -\frac{a_{12}}{a_{11}} x_2^{(k)} - \cdots - \frac{a_{1n}}{a_{11}} x_n^{(k)} + \frac{b_1}{a_{11}}, \\ x_2^{(k+1)} = -\frac{a_{21}}{a_{22}} x_1^{(k+1)} - \cdots - \frac{a_{2n}}{a_{22}} x_n^{(k)} + \frac{b_2}{a_{22}}, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x_n^{(k+1)} = -\frac{a_{n1}}{a_{nn}} x_1^{(k+1)} - \frac{a_{n2}}{a_{nn}} x_2^{(k+1)} - \cdots + \frac{b_n}{a_{nn}} \end{cases}$$

Перепишем его в матричной форме, используя те же обозначения, что и в методе Якоби:

$$x^{(k+1)} = D^{-1}(-Lx^{(k+1)} - Rx^{(k)} + b).$$

Чтобы свести к методу простой итерации, приведем подобные при $x^{(k+1)}$:

$$Dx^{(k+1)} + Lx^{(k+1)} = -Rx^{(k)} + b,\ x^{(k+1)} =$$

$$= -(L+D)^{-1}Rx^{(k)} + (L+D)^{-1}b,$$

откуда матрица B в методе простой итерации определяется соотношением

$$B = -(L+D)^{-1}R$$

Теорема 8. *Пусть $a_{ii} \neq 0$ для всех $i = 1, \ldots, n$. Тогда для сходимости метода Гаусса-Зейделя необходимо и достаточно, чтобы все корни уравнения*

$$det(\lambda L + \lambda D + R) = 0$$

удовлетворяли соотношению $|\lambda| < 1$.

Доказательство.

Согласно доказанному критерию сходимости метода простой итерации, метод сходится тогда и только тогда, когда все собственные числа матрицы B по модулю меньше единицы. Выпишем характеристическое уравнение, учитывая конкретный вид матрицы в методе Гаусса-Зейделя и проделаем преобразования

$$0 = det(\lambda E - B) = det(\lambda E + (L+D)^{-1}R) =$$

$$= det((D+L)^{-1})det(\lambda(L+D) + R),$$

откуда следует заключение теоремы.

Без доказательства приведем два достаточных условия сходимости метода Гаусса-Зейделя.

Теорема 9. *Если исходная матрица A имеет диагональное преобладание, то метод Гаусса-Зейделя сходится.*

Теорема 10. *Если исходная матрица A симметричная и положительно определена, то метод Гаусса-Зейделя сходится.*

4.9. Неустранимая погрешность при решении линейных систем. Обусловленность матриц

Рассмотрим примеры.

Пример 1. Пусть решается двумерная система

$$\begin{cases} x_1 + x_2 = 1 \\ -x_1 + x_2 = 1 \end{cases}$$

у которой точное решение $x_1 = 0, x_2 = 1$. Однако в коэффициентах системы и у чисел в правой части может быть (и реально практически всегда бывает) неустранимая погрешность. Пусть для простоты такая погрешность есть только в неоднородности, т.е. решается система

$$\begin{cases} x_1 + x_2 = 1 + \Delta b_1 \\ -x_1 + x_2 = 1 + \Delta b_2, \end{cases}$$

где $|\Delta b_1| \leq \varepsilon$, $|\Delta b_2| \leq \varepsilon$. Возникает вопрос: насколько точное решение близко к приближенному, обусловленному такой погрешностью, если величина ε мала.

Геометрическая иллюстрация приведена на рисунке, который показывает, что если ε мало, то приближенное решение близко к точному. Однако это не всегда так.

Пример 2. Пусть решается двумерная система

$$\begin{cases} 0.2x_1 + x_2 = 1 \\ x_2 = 1 \end{cases}$$

у которой точное решение $x_1 = 0$, $x_2 = 1$ и рассмотрим систему с погрешностью в неоднородности

$$\begin{cases} 0.2x_1 + x_2 = 1 + \Delta b_1 \\ x_2 = 1 + \Delta b_2, \end{cases}$$

где $|\Delta b_1| \leq \varepsilon$, $|\Delta b_2| \leq \varepsilon$. Геометрическая иллюстрация показывает, что даже если ε мало, то приближенное решение может значительно отличаться от к точного. Это пример "плохой", плохо обусловленой системы.

Перейдем к определениям.

Рассмотрим систему с точными коэффициентами и неоднородностью

$$Ax = b$$

и систему с погрешностью в неоднородности

$$A(x + \Delta x) = b + \Delta b$$

Ставится вопрос об оценке абсолютной и относительной погрешности решения в зависимости от абсолютной и относительной погрешности в неоднородности.

Вычитая из второй системы первую, получаем

$$A\Delta x = \Delta b,$$

откуда

$$\Delta x = A^{-1}\Delta b, \ \|\Delta x\| \leq \|A^{-1}\|\|\Delta b\|$$

С другой стороны, из неравенства

$$\|A\|\|x\| \geq \|Ax\| = \|b\|$$

получаем

$$\frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|}$$

Таким образом,

$$\frac{\|\Delta x\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|\Delta b\|}{\|b\|}.$$

Это неравенство показывает, во сколько раз может возрасти неустранимая погрешность (относительная) при решении линейных систем.

Числом обусловленности матрицы назовем

$$condA = \|A\|\|A^{-1}\|$$

С учетом этого определения оценка перепишется так:

$$\frac{\|\Delta x\|}{\|x\|} \leq condA\frac{\|\Delta b\|}{\|b\|}.$$

В случае, когда имеется погрешность не только в неоднородности, но и в коэффициентах системы, т.е. когда решается система

$$(A + \Delta A)(x + \Delta x) = b + \Delta b,$$

имеет место оценка

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{condA(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|})}{1 - condA\frac{\|\Delta A\|}{\|A\|}}.$$

Эту оценку приводим без доказательства.

Отметим некоторые свойства числа обусловленности.

1).

$$condA \geq 1$$

В самом деле, так как $A^{-1}A = E$, то $\|A^{-1}\|\|A\| \geq \|A^{-1}A\| = \|E\| = 1$.

Это неравенство показывает, что хорошо обусловленная матрица имеет число обусловленности близкое к единице. Так в примере 1 в поколонной норме $cond_1 A = 2$, а в примере 2 $cond_1 A = 12$ (посчитать), что приводит к большому влиянию неустранимой погрешности в примере 2.

2).

$$cond(\lambda A) = \lambda condA$$

Проверяется по определению.

3). Если матрица A симметричная, то

$$cond_2 A = \frac{\max|\lambda(A)|}{\min|\lambda(A)}$$

Доказательство вытекает из того, что для симметричной матрицы евклидова норма равна максимальному по модулю собственному числу.

5. ЧИСЛЕННОЕ РЕШЕНИЕ СИСТЕМ НЕЛИНЕЙНЫХ УРАВНЕНИЙ

5.1. Постановка задачи и предварительные сведения

Рассмотрим систему уравнений

$$\begin{cases} f_1(x_1, x_2, \cdots, x_n) = 0, \\ f_2(x_1, x_2, \cdots, x_n) = 0, \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ f_n(x_1, x_2, \cdots, x_n) = 0. \end{cases}$$

Эту систему можно записать в векторном виде

$$F(X) = 0,$$

где вектор $X = (x_1, x_2, \cdots, x_n) \in R^n$, $F(X)$ – векторная функция векторного аргумента.

Как и в случае одного уравнения, задача состоит из двух этапов: отделение корней - нахождение такой области в R^n, где существует и единственен корень $\xi \in R^n$, и уточнение корней – организация последовательности, сходящейся к корню $X^k \to \xi$.

Теория численного решения системы уравнений во многом подобна теории численного решения одного уравнения, главное отличие состоит в определении производной векторной функции по векторному аргументу.

Существуют разные определении производной векторной функции по векторному аргументу. Самыми распространенными являются производные по Фреше и по Гато.

Функция $F(X)$ называется *дифференцируемой по Фреше* в точке X, если найдется матрица P, такая, что для любого $Y \in R^n$ выполняется

$$\|F(Y) - F(X) - P(Y - X)\| = o(\|(Y - X)\|),$$

где $o(\|(Y-X)\|)$ – величина более высокого порядка, чем $\|(Y-X)\|$.

Функция $F(X)$ называется *дифференцируемой по направлению* в точке X, если для любого $Z \in R^n$ найдется матрица P, такая, что выполняется

$$\lim_{t\to\infty} \frac{F(X+tZ)-F(X)}{t} = PZ$$

Матрица P называется *производной* функции $F(X)$ в точке X.

Если матрица P не зависит от направления Z, то функцию назовем *дифференцируемой по Гато*. Заметим, что из дифференцируемости по Фреше следует дифференцируемость по Гато. В случае дифференцируемости по Гато матрицу P можно сформировать, взяв в качестве направлений $Z_j = (0,0,\ldots,0,1,0,\ldots,0)$ – j-ый единичный орт, тогда в качестве производной функции $F(X)$ в точке X выступает матрица Якоби

$$F'(X) = \begin{pmatrix} \frac{\partial f_1(X)}{\partial x_1} & \ldots & \frac{\partial f_1(X)}{\partial x_n} \\ . & \ldots & . \\ \frac{\partial f_n(X)}{\partial x_1} & \ldots & \frac{\partial f_n(X)}{\partial x_n} \end{pmatrix}$$

5.2. Метод Ньютона для решения систем нелинейных уравнений

Напомним, что для решения одномерного уравнения $f(x)=0$, метод Ньютона записывался так

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

Многомерный аналог можно получить, взяв линейную часть в тейлоровском разложении

$$0 = F(X) = F(X^{(k)}) + F'(X^{(k)})(X - X^{(k)}) + \cdots$$

тогда

$$X^{(k+1)} = X^{(k)} - (F'(X^{(k)}))^{-1} F(X^{(k)})$$

Эти формулы содержат трудоемкую операцию взятия обратной матрицы, поэтому преобразуем их, перейдя к приращениям

$$\Delta X^{(k)} = X^{(k+1)} - X^{(k)},$$

и умножив равенство слева на $F'(X^{(k)})$. Тогда метод перепишется в виде

$$F'(X^{(k)})\Delta X^{(k)} = -F(X^{(k)}).$$

Это метод Ньютона-Рафсона, называемый также методом Ньютона-Канторовича. Часто его также называют методом лианеризации, так как на каждом шаге он представляет собой задачу решения линейных систем относительно $\Delta X^{(k)}$, сводя, таким образом, задачу о решении нелинейной системы к решению последовательности линейных систем.

5.3. Метод простой итерации для решения систем нелинейных уравнений

Перейдем от исходной системы к системе вида

$$\begin{cases} x_1 = \varphi_1(x_1, x_2, \cdots, x_n), \\ x_2 = \varphi_2(x_1, x_2, \cdots, x_n), \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ x_3 = \varphi_n(x_1, x_2, \cdots, x_n), \end{cases}$$

или в векторном виде

$$X = \Phi(X).$$

Методом простой итерации для нелинейных систем назовем алгоритм

$$X^{(k+1)} = \Phi(X^{(k)}).$$

Как и в одномерном случае, метод простой итерации является базовым, т.к. все другие итерационные методы можно представить как метод простой итерации и тем самым исследовать вопрос о сходимости. Справедливо утверждение

Теорема 11. *Пусть функция* $\Phi(X)$ *непрерывно дифференцируема в окрестности решения* ξ *системы, причем выполняется* $\|\Phi'(\xi)\| < 1$ *для подчиненной нормы. Тогда существует такая окрестность* $U(\xi)$ *вектора* ξ*, что начальное приближение* $X^{(0)}$ *принадлежит этой окрестности, то метод простой итерации сходится к* ξ*.*

Доказательство.

Рассмотрим разность скалярной функции от векторного аргумента $\varphi_i(Y) - \varphi_i(X)$.

Сведем разность значений функции векторного аргумента к разности значений функции скалярного аргумента. Введем вспомогательную векторную функцию скалярного аргумента:

$$X(t) = X + t(Y - X),$$

тогда

$$X(0) = X, X(1) = Y,$$

$$\varphi_i(Y) - \varphi_i(X) = \varphi_i(X(1)) - \varphi_i(X(0)) = \psi_i(1) - \psi_i(0).$$

Здесь $\psi_i(t) = \varphi_i(X(t))$ – сложная функция.

Применим к функции $\psi_i(t)$ формулу конечных приращений Лагранжа

$$\psi_i(1) - \psi_i(0) = \psi_i'(c_i)(1 - 0),\ c_i \in [0; 1]$$

Вычисляя производную функции $\psi_i(t)$, получаем

$$\psi_i(1) - \psi_i(0) = \sum_{j=1}^{n} \frac{\partial \varphi_i}{\partial x_j}(X(c_i))(y_j - x_j)$$

Возвращаясь к векторной функции векторного аргумента, получаем

$$\Phi(Y) - \Phi(X) = B(c_1, c_2, \cdots, c_n)(Y - X),$$

где

$$B = B(c_1, c_2, \cdots, c_n) = \begin{pmatrix} \frac{\partial \varphi_1(X(c_1))}{\partial x_1} & \cdots & \frac{\partial \varphi_1(X(c_1))}{\partial x_n} \\ . & \cdots & . \\ \frac{\partial \varphi_n(X(c_n))}{\partial x_1} & \cdots & \frac{\partial \varphi_n(X(c_n))}{\partial x_n} \end{pmatrix}$$

Пусть теперь ξ – решение системы и в силу непрерывности производной выполняется $\|\Phi'(X)\| \leq q < 1$ в некоторой окрестности $U(\xi)$ вектора ξ. В силу непрерывности частных производных выполняется $\|B\| \leq q < 1$, если X и Y принадлежат этой окрестности. Тогда

$$\|\Phi(Y) - \Phi(X)\| \leq q\|Y - X\|$$

Сузим $U(\xi)$ до его подмножества – некоторого шара в той метрике, в которой рассматривается подчиненная норма, с центром в ξ, обозначать будем также.

Докажем, что если некоторое приближение $X^k \in U(\xi)$, то $X^{k+1} \in U(\xi)$. В самом деле

$$\|X^{k+1} - \xi\| \leq q\|X^k - \xi\| \leq \|X^k - \xi\|$$

Таким образом, если взять начальное приближение X^0 из этой окрестности, то все последующие приближения также будут в этой окрестности. Имеем

$$\|X^k - \xi\| \leq q\|X^{k-1} - \xi\| \leq \cdots \leq q^k\|X^0 - \xi\|$$

откуда следует сходимость $X^k \to \xi$ при $k \to \infty$.

6. ИНТЕРПОЛЯЦИЯ

Теория интерполяции в численных методах важна как базовый аппарат для решения таких задач, как численное дифференцирование, численное интегрирование, численное решение дифференциальных уравнений.

6.1. Постановка задачи

Даны числа $x_0, x_1, \ldots, x_n$, называемые узлами интерполяции (пока для простоты попарно различные), а также числа $f_0, f_1, \ldots, f_n$, называемые значениями в узлах. Требуется найти функцию $f(x)$, удовлетворяющую интерполяционным условиям

$$f(x_i) = f_i,\ i = 0, 1, \ldots, n.$$

Конечно, если не оговорить класс функций, то задача интерполяции может иметь неединственное решение. Будем выбирать функцию $f(x)$ в классе линейных комбинаций некоторых функций $\varphi_i(x),\ i = 0, 1, \ldots, n$:

$$f(x) = C_0\varphi_0(x) + C_1\varphi_1(x) + \cdots + C_n\varphi_n(x).$$

Тогда задача интерполяции сводится к решению линейной системы относительно коэффициентов $C_0, C_1, \ldots, C_n$:

$$\begin{cases} C_0\varphi_0(x_0) + C_1\varphi_1(x_0) + \cdots + C_n\varphi_n(x_0) = f_0 \\ C_0\varphi_0(x_1) + C_1\varphi_1(x_1) + \cdots + C_n\varphi_n(x_1) = f_1 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ C_0\varphi_0(x_n) + C_1\varphi_1(x_n) + \cdots + C_n\varphi_n(x_n) = f_n \end{cases}$$

Система однозначно разрешима тогда и только тогда, когда её главный определитель отличен от нуля. Это свойство называется *линейной незави-*

симостью на системе узлов. Отметим, что одна и та же система функций может быть на одной системе узлов зависимой, а на другой – независимой.

Пример: Пусть $\varphi_0 \equiv 1$, $\varphi_1 = x^2$. Эта система функций независима на системе узлов $x_0 = 0$, $x_1 = 1$, но зависима на системе узлов $x_0 = -1$, $x_1 = 1$.

Особую роль играют системы функций, линейно независимые на любой системе узлов. Такие системы функций называются it чебышевскими. Докажем, что система функций $\varphi_0 \equiv 1$, $\varphi_1 = x$, $\varphi_2 = x^2, \ldots$, $\varphi_n = x^n$ чебышевская. В самом деле, для любой системы попарно различных узлов $x_0, x_1, \ldots, x_n$, нужно проверить отличие от нуля определителя

$$\begin{vmatrix} 1 & x_0 & x_0^2 & ... & x_0^n \\ 1 & x_1 & x_1^2 & ... & x_1^n \\ . & . & . & ... & . \\ 1 & x_n & x_n^2 & ... & x_n^n \end{vmatrix}$$

Но это определитель Ван-Дер-Монда, который равен $\prod\limits_{i \neq j}(x_i - x_j)$, и, так как узлы попарно различны, то определитель отличен от нуля. Система чебышевская.

В дальнейшем мы будем рассматривать только эту систему, рассматривая таким образом задачу интерполяции в классе многочленов $n-$ой степени. В силу чебышевости этот многочлен, $f(x) = L_n(x) = C_0 + C_1 x + C_2 x^2 + \cdots + C_n x^n$, обозначаемый в честь Лагранжа, единственным образом решает задачу интерполяции. Существуют две формы его конструктивного построения, применяемые в различных ситуациях.

6.2. Интерполяционный многочлен в форме Лагранжа

Рассмотрим сначала вспомогательную задачу интерполяции, когда все значения интерполяционного многочлена в узлах равны 0, за исключением i-го узла, где значение равно 1. Решение этой задачи дается формулой

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

В самом деле при всяком $i = 0, 1, \cdots, n$, функция $l_i(x)$ является многочленом $n-$ой степени; простой подстановкой получаем, что $l_i(x_i) = 1$, $l_i(x_j) = 0, j \neq i$, а в силу единственности других многочленов этой степени, решающих вспомогательную задачу нет.

Теперь рассмотрим исходную задачу интерполяции, её решение выражается формулой

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

откуда, подставляя выражение для $l_i(x)$, получаем формулу Лагранжа

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

Замечание. Интерполяционный многочлен в форме Лагранжа удобно применять, когда узлы интерполяции зафиксированы, а меняются значения функции (или функций).

6.3. Погрешность интерполяционного многочлена Лагранжа

Будем предполагать, как и всюду в курсе, что выполняются требуемые для действий условия гладкости, в данном случае, что интерполируемая функция $f(x)$ $n+1$-раз непрерывно дифференцируема.

Обозначим погрешность интерполяции $R_n(x) = f(x) - L_n(x)$. По постановке задачи погрешность интерполяции в узлах $x_0, x_1, \ldots, x_n$ равна нулю. Но такие же корни имеет функция

$$\omega_n(x) = (x - x_0)(x - x_1) \cdots (x - x_n).$$

Представим погрешность в виде

$$R_n(x) = K(x)\omega_n(x)$$

и попробуем найти функцию $K(x)$. Зафиксируем число $\hat{x} \neq x_i, i = 0, \ldots, n$ и введем вспомогательную функцию

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

Функция $F(x)$ имеет по крайней мере $n+2$ различных корня – узлы интерполяции $x_0, x_1, \ldots, x_n$, и число $\hat{x}$. По теореме Ролля, её первая производная $F'(x)$ имеет по крайней мере $n+1$ различный корень и т.д. $n+1$-производная имеет по крайней мере один корень, обозначим его ξ. Таким образом

$$F^{(n+1)}(\xi) = 0.$$

Но

$$F^{(n+1)}(x) = f^{(n+1)}(x) - K(\hat{x})(n+1)!,$$

поэтому

$$f^{(n+1)}(\xi) - K(\hat{x})(n+1)! = 0,$$

откуда

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

Подставим $K(\hat{x})$ в формулу для погрешности интерполяции, расфиксировав $\hat{x}$, получаем

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_n(x), \ \xi \in [x_0, x_1, \cdots, x_n, x].$$

Здесь и в дальнейшем $[x_0, x_1, \cdots, x_n, x]$ означает наименьший отрезок, содержащий узлы интерполяции $x_0, x_1, \cdots, x_n$ и точку x, в которой происходит интерполяция.

6.4. Разделенные разности и интерполяционный многочлен в форме Ньютона

Для того, чтобы построить интерполяционный многочлен в другой форме, более удобной в некоторых ситуациях, необходимо познакомиться с элементами теории разделенных разностей, созданной Ньютоном.

Разделённые разности вводятся реккурентно. Пусть даны попарно различные узлы $x_0, x_1, \ldots, x_n$, и значения в узлах $f_0, f_1, \ldots, f_n$.

Разделенными разностями нулевого порядка назовем значения в узлах $f(x_i) = f_i$.

Разделенной разностью первого порядка по узлам x_i, x_j, назовем величину

$$f(x_i, x_j) = \frac{f(x_i) - f(x_j)}{x_i - x_j}$$

Далее разделенные разности k-го порядка вводятся через разделенные разности $k-1$-го порядка.

Разделенной разностью k-го порядка по узлам $x_0, x_1, \ldots, x_k$, назовем величину

$$f(x_0, x_1, \ldots, x_k) = \frac{f(x_1, x_2, \ldots, x_k) - f(x_0, x_1, \ldots, x_{k-1})}{x_k - x_0}$$

Среди свойств разделенных разностей отметим следующее представление

Лемма.

$$f(x_0, x_1, \ldots, x_n) = \sum_{k=0}^{n} \frac{f(x_k)}{\prod\limits_{i=0, i\neq k}^{n} (x_k - x_i)}$$

Доказательство проведем индукцией по n.

База индукции при $n = 1$:

$$f(x_0, x_1) = \frac{f(x_0) - f(x_1)}{x_0 - x_1} = \frac{f(x_0)}{x_0 - x_1} + \frac{f(x_1)}{x_1 - x_0}$$

Шаг индукции. Предположим, что представление справедливо при $n-1$, докажем представление при n. По определению разделенной разности n-го порядка и по индуктивному предположению получаем, приводя подобные

$$f(x_0, x_1, \ldots, x_n) = \frac{f(x_1, x_2, \ldots, x_n) - f(x_0, x_1, \ldots, x_{n-1})}{x_n - x_0} =$$

$$= \sum_{k=0}^{n-1} \frac{f(x_k)}{(x_0 - x_n) \prod\limits_{i=0, i\neq k}^{n-1} (x_k - x_i)} - \sum_{k=1}^{n} \frac{f(x_k)}{(x_0 - x_n) \prod\limits_{i=1, i\neq k}^{n} (x_k - x_i)} =$$

$$= \frac{f(x_0)}{\prod\limits_{i=1}^{n}(x_0 - x_i)} + \frac{f(x_n)}{\prod\limits_{i=0}^{n-1}(x_n - x_i)} +$$

$$+ \sum_{k=1}^{n-1} \frac{f(x_k)}{(x_0 - x_n) \prod\limits_{i=1, i\neq k}^{n-1} (x_k - x_i)} \left[\frac{1}{x_k - x_0} - \frac{1}{x_k - x_n}\right] =$$

$$= \frac{f(x_0)}{\prod\limits_{i=1}^{n}(x_0 - x_i)} + \frac{f(x_n)}{\prod\limits_{i=0}^{n-1}(x_n - x_i)} + \sum_{k=1}^{n-1} \frac{f(x_k)}{\prod\limits_{i=0, i\neq k}^{n} (x_k - x_i)} =$$

$$= \sum_{k=0}^{n} \frac{f(x_k)}{\prod\limits_{i=0, i\neq k}^{n} (x_k - x_i)}$$

Лемма доказана.

Замечание. Из леммы, в частности, следует, что разделенная разность является симметричной функцией своих аргументов, т.е. порядок аргументов можно менять.

Вернемся к решению задачи интерполяции. Справедливо утверждение.

Теорема 12. *А.*

$$L_n(x) = f(x_0) + f(x_0, x_1)(x - x_0) + f(x_0, x_1, x_2)(x - x_0)(x - x_1) + \cdots +$$

$$+ f(x_0, \cdots, x_n)(x - x_0) \cdots (x - x_{n-1})$$

Б.

$$R_n(x) = f(x, x_0, \cdots, x_n)(x - x_0) \cdots (x - x_n)$$

Доказательство. Докажем сначала часть Б. Используя определение погрешности интерполяции и формулу интерполяционного многочлена в форме Лагранжа, получаем

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

$$= [\frac{f(x)}{\prod\limits_{j=0}^{n} (x - x_j)} - \sum_{i=0}^{n} \frac{f(x_i)}{x - x_i}] \prod_{j=0}^{n} (x - x_j)$$

Сравнивая выражение в квадратных скобках с утверждением леммы (считая что x дополнительный аргумент разделенной разности), получаем

$$R_n(x) = f(x, x_0, \cdots, x_n) \prod_{j=0}^{n} (x - x_j).$$

Докажем часть А. Возьмем тождество

$$L_n(x) = L_0(x) + (L_1(x) - L_0(x)) + (L_2(x) - L_1(x)) + \cdots +$$

$$+ (L_n(x) - L_{n-1}(x)),$$

и для любого $m = 1, \ldots, n$, рассмотрим разности $L_m(x) - L_{m-1}(x)$. Эти разности являются многочленом степени m, имеют m корней $x_0, x_1, \ldots, x_{m-1}$

(так как $L_m(x_i) = L_{m-1}(x_i) = f_i$ для $i = 0, 1, \ldots, m-1$), следовательно, имеет место представление

$$L_m(x) - L_{m-1}(x) = A_m(x - x_0)(x - x_1)\cdots(x - x_{m-1}.$$

Константу A_m найдем, подставляя в разности число x_m:

$$L_m(x_m) - L_{m-1}(x_m) = f(x_m) - L_{m-1}(x_m) =$$

$$= A_m(x_m - x_0)(x_m - x_1)\cdots(x_m - x_{m-1}),$$

с другой стороны из части Б теоремы вытекает, что

$$f(x_m) - L_{m-1}(x_m) = f(x_m, x_0, \cdots, x_{m-1})(x_m - x_0)(x_m - x_1)\cdots$$

$$\cdots(x_m - x_{m-1}),$$

откуда $A_m = f(x_m, x_0, \cdots, x_{m-1})$. Подставляя найденные выражения в тождество, получаем часть А теоремы.

Замечание 1. Интерполяционный многочлен в форме Ньютона удобно применять, когда интерполируемая функция зафиксирована, а узлы интерполяции добавляются для обеспечения большей точности. В этом случае коэффициенты многочлена в форме Ньютона непересчитываются. Вычисления удобно оформлять в виде таблицы разделенных разностей.

Замечание 2. Сравнивая погрешности интерполяционного многочлена в форме Лагранжа и в форме Ньютона, получаем связь между разделенными разностями и производными. Пусть узлы $x_0, x_1, \ldots, x_k$, попарно различны. Тогда

$$f(x_0, x_1, \cdots, x_k) = \frac{f^{(k)}(\xi)}{k!},\ \xi \in [x_0, x_1, \cdots, x_k]$$

6.5. Интерполяция с кратными узлами

Кратный узел означает, что в этом узле задано не только значение интерполируемой функции, но и производные до какого-либо порядка. Пусть даны узел x_1 кратности m_1, узел x_2 кратности m_2, и т.д. узел x_s кратности m_s, т.е. следующие интерполяционные данные

$$\begin{array}{cccc} x_1 & f'(x_1) & \ldots & f^{(m_1-1)}(x_1) \\ x_2 & f'(x_2) & \ldots & f^{(m_2-1)}(x_2) \\ . & . & \ldots & . \\ x_s & f'(x_s) & \ldots & f^{(m_s-1)}(x_s) \end{array}$$

Всего интерполяционных данных $m_1+m_2+\cdots+m_s = n-1$. Требуется построить многочлен $H_n(x)$ (обозначаемый в честь Эрмита), степени $n = m_1 + m_2 + \cdots + m_s - 1$, удовлетворяющий условиям

$$H_n^{(j)}(x_i) = f^{(j)}(x_i),\ i = 1, \ldots, s,\ j = 0, \ldots, m_i - 1.$$

Докажем, что эта задача имеет единственное решение. В самом деле, задача представляет собой неоднородную линейную систему относительно коэффициентов многочлена $H_n(x)$. А линейная неоднородная система имеет единственное решение тогда и только тогда, когда соответствующая однородная система

$$H_n^{(j)}(x_i) = 0,\ i = 1, \ldots, s,\ j = 0, \ldots, m_i - 1$$

имеет лишь тривиальное решение. Но последняя система означает, что многочлен n-ой степени имеет $n + 1$ корень (с учетом кратности). А это возможно, если только многочлен нулевой.

Для того, чтобы построить искомый многочлен Эрмита $H_n(x)$, рассмотрим вспомогательные конструкции.

6.6. Разделённые разности с кратными узлами

Пусть даны узлы $x_i, x_{i+1}, \ldots, x_{i+n}$, среди которых могут быть и совпадающие. Для того, чтобы определить разделенную разность по этим узлам, заменим набор узлов набором узлов $x_i^\varepsilon, x_{i+1}^\varepsilon, \ldots, x_{i+n}^\varepsilon$, который во-первых, в отличие от исходного, состоит из попарно различных узлов, а во вторых $x_{i+k}^\varepsilon \to x_{i+k}$ при $\varepsilon \to 0$. Такие наборы можно всегда организовать, например, если узлы $x_i, x_{i+1}, \ldots, x_{i+k}$ совпадают, то можно при малом ε взять $x_{i+j}^\varepsilon = x_i + j\varepsilon$ при $j = 0, 1, \cdots, k$.

Разделенной разностью по возможно совпадающим узлам назовем

$$f(x_i, x_{i+1}, \ldots, x_{i+n}) = \lim_{\varepsilon \to 0} f(x_i^\varepsilon, x_{i+1}^\varepsilon, \ldots, x_{i+n}^\varepsilon),$$

если такой предел существует.

Отметим два частных случая для конструктивного вычисления такой разделенной разности.

А. Пусть среди узлов есть хотя бы одна пара различных. Пусть это первый узел x_i и последний x_{i+n}. Тогда при фиксированном ε можно воспользоваться реккурентным определением разделенной разности (узлы различные) и сохранить при переходе к пределу реккурентное определение.

$$f(x_i, x_{i+1}, \ldots, x_{i+n}) = \lim_{\varepsilon \to 0} f(x_i^\varepsilon, x_{i+1}^\varepsilon, \ldots, x_{i+n}^\varepsilon) =$$

$$= \lim_{\varepsilon \to 0} \frac{f(x_{i+1}^\varepsilon, \ldots, x_{i+n}^\varepsilon) - f(x_i^\varepsilon, \ldots, x_{i+n-1}^\varepsilon)}{x_{i+n}^\varepsilon - x_i^\varepsilon} =$$

$$= \frac{f(x_{i+1}, \ldots, x_{i+n}) - f(x_i, \ldots, x_{i+n-1})}{x_{i+n} - x_i}$$

Б. Пусть все узлы совпадают $x_i = x_{i+1} = \ldots = x_{i+n}$. Тогда при фиксированном ε можно воспользоваться связью разделенных разностей и производных и сохранить эту связь при переходе к пределу.

$$f(x_i, x_{i+1}, \ldots, x_{i+n}) = f(x_i, x_i, \ldots, x_i) =$$

$$= \lim_{\varepsilon \to 0} f(x_i^\varepsilon, x_{i+1}^\varepsilon, \ldots, x_{i+n}^\varepsilon) = \lim_{\varepsilon \to 0} \frac{f^{(k)}(\xi^\varepsilon)}{n!} = \frac{f^{(k)}(x_i)}{n!}$$

6.7. Интерполяционный многочлен Эрмита

Рассмотрим исходную задачу интерполяции с кратными узлами и возьмем узлы с учетом кратности, введя двойную нумерацию узлов:

$$x_{1,1} = x_{1,2} = \cdots = x_{1,m_1}, x_{2,1} = x_{2,2} = \cdots = x_{2,m_2}, \cdots,$$

$$x_{s,1} = x_{s,2} = \cdots = x_{s,m_s}$$

Заменим этот набор узлов набором попарно различных узлов

$$x_{1,1}^\varepsilon = x_{1,2}^\varepsilon = \cdots = x_{1,m_1}^\varepsilon, x_{2,1}^\varepsilon = x_{2,2}^\varepsilon = \cdots = x_{2,m_2}^\varepsilon, \cdots,$$

$$x_{s,1}^\varepsilon = x_{s,2}^\varepsilon = \cdots = x_{s,m_s}^\varepsilon$$

и построим по ним интерполяционный многочлен степени n в форме Ньютона. Переходя к пределу при $\varepsilon \to 0$, получаем

$$H_n(x) = f(x_1) + f'(x_1)(x - x_1) + \cdots + \frac{f^{(m_1-1)}(x_1)}{(m_1-1)!}(x - x_1)^{m_1-1} +$$

$$+ f(x_1, x_1, \cdots, x_1, x_2)(x - x_1)^{m_1} + \cdots +$$

$$+ f(x_1, x_1, \cdots, x_1, \cdots, x_s)(x - x_1)^{m_1} \cdots (x - x_s)^{m_s - 1}$$

При этом погрешность выражается формулой

$$R_n(x) = f(x, x_1, x_1, \cdots, x_1, \cdots, x_s)(x - x_1)^{m_1} \cdots (x - x_s)^{m_s}$$

Заметим, что формула Эрмита с одной стороны обобщает формулу Ньютона с её аппаратом разделенных разностей, а с другой стороны формулу Тейлора с её аппаратом производных.

6.8. Дополнительные свойства разделенных разностей

Установим связь между разделенными разностями и производными в общем случае. Выше такая связь была установлена в двух частных случаях: когда узлы попарно различны и когда все узлы совпадают. По определению

$$f(x_i, x_{i+1}, \ldots, x_{i+n}) = \lim_{\varepsilon \to 0} f(x_i^\varepsilon, x_{i+1}^\varepsilon, \ldots, x_{i+n}^\varepsilon).$$

Так как узлы $x_i^\varepsilon, x_{i+1}^\varepsilon, \ldots, x_{i+n}^\varepsilon$ попарно различные, то для разделенной разности по этим узлам можно воспользоваться связью с производной и получаем

$$f(x_i, x_{i+1}, \ldots, x_{i+n}) = \lim_{\varepsilon \to 0} \frac{f^{(n)}(\xi^\varepsilon)}{n!},\ \xi^\varepsilon \in [x_i^\varepsilon, x_{i+1}^\varepsilon, \ldots, x_{i+n}^\varepsilon].$$

Последовательность ξ^ε ограничена, из неё можно выделить сходящуюся подпоследовательность $\xi^\varepsilon \to \xi$, $\xi \in [x_i, x_{i+1}, \ldots, x_{i+n}]$. Таким образом

$$f(x_i, x_{i+1}, \ldots, x_{i+n}) = \frac{f^{(n)}(\xi)}{n!},\ \xi \in [x_i, x_{i+1}, \ldots, x_{i+n}]$$

в случае произвольных узлов.

Для дальнейшего нам потребуется уметь дифференцировать разделенную разность по одному из узлов. Справедлива формула

$$\frac{d}{dx} f(x, x_i, x_{i+1}, \ldots, x_{i+n}) = f(x, x, x_i, x_{i+1}, \ldots, x_{i+n}).$$

В самом деле, по определению производной и индуктивному определению разделенной разности

$$\frac{d}{dx} f(x, x_i, x_{i+1}, \ldots, x_{i+n}) =$$

$$= \lim_{\Delta x \to 0} \frac{f(x + \Delta x, x_i, x_{i+1}, \ldots, x_{i+n}) - f(x, x_i, x_{i+1}, \ldots, x_{i+n})}{\Delta x} =$$

$$= \lim_{\Delta x \to 0} f(x + \Delta x, x, x_i, x_{i+1}, \ldots, x_{i+n}) = f(x, x, x_i, x_{i+1}, \ldots, x_{i+n}).$$

Аналогичным образом проверяется формула

$$\frac{d^2}{dx^2} f(x, x_i, x_{i+1}, \ldots, x_{i+n}) = 2f(x, x, x, x_i, x_{i+1}, \ldots, x_{i+n}).$$

7. ЧИСЛЕННОЕ ДИФФЕРЕНЦИРОВАНИЕ

Даны узлы $x_0, x_1, \ldots, x_n$, и значения функции в узлах $f_0, f_1, \ldots, f_n$, требуется найти значения производных функции в некоторых точках.

7.1. Общий подход к численному дифференцированию

Общий подход состоит в том, что строится интерполяционный многочлен $L_n(x)$ и он дифференцируется, при этом погрешности численного дифференцирования R_∂ равна производной погрешности интерполяции:

$$f'(x) = L'_n(x) + R'_n(x),\ f'(x) \approx L'_n(x),\ R_\partial = R'_n(x).$$

7.2. Численное дифференцирование по двум узлам

Заданы узлы x_0, x_1, (в дальнейшем будем считать, что $h = x_1 - x_0 > 0$) и значения f_0, f_1 в узлах. Выпишем интерполяционный многочлен, например, в форме Ньютона, и его продифференцируем.

$$L_1(x) = f(x_0) + f(x_0, x_1)(x - x_0),\ L'_1(x) = f(x_0, x_1) = \frac{f_1 - f_0}{h},$$

тогда формула численного дифференцирования по двум узлам будет

$$f'(x) \approx \frac{f_1 - f_0}{h}$$

Погрешность интерполяции

$$R_1(x) = f(x, x_0, x_1)(x - x_0)(x - x_1),$$

дифференцируя погрешность, в том числе используя правило дифференцирования разделенной разности по узлу, получаем

$$R_\partial = R'_1(x) = f(x, x, x_0, x_1)(x - x_0)(x - x_1) +$$

$$+f(x,x_0,x_1)(2x-x_0-x_1).$$

Погрешность дифференцирования зависит от x, в отличие от самой формулы. Рассмотрим наиболее важные частные случаи.

А. Дифференцирование на левый край $x=x_0$. Тогда, используя связь разделенной разности и производной, получаем

$$R_\partial = -f(x_0,x_0,x_1)h = -\frac{f''(\xi)}{2}h,\ \xi\in[x_0,x_1]$$

Аналогичным образом, при дифференцировании на правый край $x=x_1$, получаем

$$R_\partial = -f(x_0,x_0,x_1)h = \frac{f''(\xi)}{2}h,\ \xi\in[x_0,x_1]$$

Обычно при дифференцировании на левый или правый край применяют оценку погрешности

$$|R_\partial| \le \frac{M_2}{2}h,\ M_2 = \max_{x\in[x_0,x_1]}|f''(x)|$$

Б. Дифференцирование на середину $x=\frac{x_0+x_1}{2}$. Используя связь разделенной разности и производной, получаем

$$R_\partial = -f(\frac{x_0+x_1}{2},\frac{x_0+x_1}{2},x_0,x_1)\frac{h^2}{4} = -f'''(\xi)\frac{h^2}{24},\ \xi\in[x_0,x_1],$$

и оценку

$$|R_\partial| \le \frac{M_3}{24}h^2,\ M_3 = \max_{x\in[x_0,x_1]}|f'''(x)|$$

Сравнивая погрешности численного дифференцирования на середину и край, заметим, что порядок малости по h на середину второй, а на край только первый.

7.3. Численное дифференцирование по трем узлам

Пусть даны три равноотстоящих узла x_0,x_1,x_2 ($x_1-x_0=x_2-x_1=h>0$) и значения f_0,f_1,f_2 в узлах. Требуется построить формулы чис-

ленного дифференцирования для первой и второй производной и оценить погрешности.

Выпишем интерполяционный многочлен, например, в форме Ньютона, и его продифференцируем два раза

$$L_2(x) = f(x_0) + f(x_0, x_1)(x - x_0) + f(x_0, x_1, x_2)(x - x_0)(x - x_1),$$

$$L_2'(x) = f(x_0, x_1) + f(x_0, x_1, x_2)(2x - x_0 - x_1),$$

$$L_2''(x) = 2f(x_0, x_1, x_2).$$

Распишем разделенную разность второго порядка

$$f(x_0, x_1, x_2) = \frac{f(x_1, x_2) - f(x_0, x_1)}{2h} = \frac{\frac{f_2 - f_1}{h} - \frac{f_1 - f_0}{h}}{2h} =$$

$$= \frac{f_0 - 2f_1 + f_2}{2h^2}$$

тогда формулы численного дифференцирования по трем узлам будут

$$f'(x) \approx \frac{f_1 - f_0}{h} + \frac{f_0 - 2f_1 + f_2}{2h^2}(2x - x_0 - x_1),$$

$$f''(x) \approx \frac{f_0 - 2f_1 + f_2}{h^2}$$

Подставляя различные x в формулу для первой производной мы можем получить несколько вариантов, например, формула численного дифференцирования на левый край ($x = x_0$) по трем равноотстоящим узлам выглядит так:

$$f'(x_0) \approx \frac{-3f_0 + 4f_1 - f_2}{2h}$$

Но наибольшее значение имеет формула численного дифференцирования для второй производной по трем равноотстоящим узлам, которая не зависит от x, это вообще одна из самых распространенных формул в современной математике.

Изучим вопрос о погрешности, особенно формулы для второй производной. Погрешность интерполяции

$$R_2(x) = f(x, x_0, x_1, x_2)(x - x_0)(x - x_1)(x - x_2)$$

и вычисление производных становится очень громоздким. Чтобы немного сократить записи, введем обозначения $\alpha_i = x - x_i,\ i = 0, 1, 2$, тогда

$$R_2(x) = f(x, x_0, x_1, x_2)\alpha_0\alpha_1\alpha_2$$

$$R_2'(x) = f(x, x, x_0, x_1, x_2)\alpha_0\alpha_1\alpha_2 +$$

$$+f(x, x_0, x_1, x_2)[\alpha_1\alpha_2 + \alpha_0\alpha_2 + \alpha_0\alpha_1]$$

Отсюда можно получать оценки погрешности при разных x для формул численного дифференцирования для первой производной. Но мы это делать не будем, а найдем вторую производную

$$R_2''(x) = 2f(x, x, x, x_0, x_1, x_2)\alpha_0\alpha_1\alpha_2 +$$

$$+2f(x, x, x_0, x_1, x_2)[\alpha_1\alpha_2 + \alpha_0\alpha_2 + \alpha_0\alpha_1] +$$

$$+2f(x, x_0, x_1, x_2)(\alpha_0 + \alpha_1 + \alpha_2)$$

Рассмотрим наиболее важный случай – формулу дифференцирования на середину ($x = x_1$) для второй производной. Тогда $\alpha_1 = 0$, $\alpha_0 = h$, $\alpha_2 = -h$, и в формуле для погрешности остается только одно слагаемое

$$R_\partial = R_2''(x_1) = 2f(x_1, x_1, x_0, x_1, x_2)(-h^2)$$

Выражая разделенную разность четвертого порядка через соответствующую производную, получаем погрешность

$$R_\partial == -f''''(\xi)\frac{h^2}{12},\ \xi \in [x_0, x_2]$$

и оценку погрешности

$$|R_\partial| \le \frac{M_4}{12}h^2,\ M_4 = \max_{x\in[x_0,x_2]} |f''''(x)|$$

7.4. Метод неопределенных коэффициентов

Приведем простой способ вывода формул численного дифференцирования, этот способ можно успешно применять и в других разделах численных методов. Так как интерполяционный многочлен линеен относительно значений интерполируемой функции, а операция дифференцирования также линейна, то и формула численного дифференцирования линейна относительно заданных значений функции. Это дает возможность искать формулу в виде линейной комбинации заданных значений функции. Коэффициенты линейной комбинации можно получить, подставляя в формулу многочлены, начиная с низших степеней, так как интерполяционный многочлен совпадает с интерполируемой функцией, которая сама является многочленом не более высокой степени, а дифференцирование сохраняет равенство.

Проиллюстрируем метод примером. Пусть в трех равноотстоящих узлах x_0, x_1, x_2 ($x_1 - x_0 = x_2 - x_1 = h > 0$) заданы значения f_0, f_1, f_2 функции $f(x)$. Требуется построить формулу для $f'(x_0)$.

Запишем искомую формулу в виде

$$f'(x_0) \approx Af_0 + Bf_1 + Cf_2.$$

Приближенное равенство становится точным, если $f(x)$ является многочленом не выше второй степени. Последовательно подставим в равенство $f(x) \equiv 1$, $f(x) = x - x_0$, $f(x) = (x - x_0)^2$, получим систему уравнений

$$\begin{cases} 0 = A + B + C \\ 1 = Bh + 2CH \\ 0 = BH^2 + 4Ch^2, \end{cases}$$

Решая эту систему получаем $C = -1/(2h)$, $B = 2/h$, $A = -3/(2h)$.

Подставляя коэффициенты, получаем формулу

$$f'(x_0) \approx \frac{-3f_0 + 4f_1 - f_2}{2h},$$

уже выведенную нами ранее путем построения интерполяционного многочлена и его последующего дифференцирования.

Отметим, что метод неопределенных коэффициентов очень удобен для вывода формул, но не дает оценку погрешностей.

7.5. Неустранимая погрешность при численном дифференцировании

На примере формулы численного дифференцирования по двум узлам на левый край рассмотрим погрешность с учетом неустранимой.

Пусть вместо точных значений функции f_0, f_1 заданы их приближенные значения f_0^*, f_1^* с погрешностью A_f. Оценим погрешность формулы

$$f'(x_0) \approx \frac{f_1^* - f_0^*}{h}$$

С учетом погрешности метода и влияния неустранимой погрешности имеем

$$|f'(x_0) - \frac{f_1^* - f_0^*}{h}| \le |f'(x_0) - \frac{f_1 - f_0}{h}| + |\frac{f_1 - f_0}{h} - \frac{f_1^* - f_0^*}{h}| \le$$

$$\le \frac{M_2}{2}h + \frac{2A_f}{h}$$

Таким образом, полная погрешность этой формулы численного дифференцирования оценивается (без учета влияния вычислительной погрешности) величиной

$$A_{пол} = \frac{M_2}{2}h + \frac{2A_f}{h}$$

Эта оценка показывает, что при уменьшении величины шага погрешность сначала уменьшается, а затем увеличивается. Причина этого явления состоит в том, что операция численного дифференцирования относится к

числу *некорректных задач*, т.е. задач, в которых малые изменения в исходных данных могут дать большие изменения в результате. Это явление характерно не только для данной формулы, но и для других формул численного дифференцирования.

7.6. Выбор оптимального шага при численном дифференцировании

Как показывает формула полной погрешности численного дифференцирования по двум узлам на левый край, существует оптимальный шаг $h_{опт}$ при котором полная погрешность минимальна. Решая средствами математического анализа задачу на экстремум, получаем

$$A'_{пол}(h_{опт}) = \frac{M_2}{2} - \frac{2A_f}{h^2} = 0,\ h_{опт} = 2\sqrt{\frac{A_f}{M_2}},$$

$$A_{пол}(h_{опт}) = 2\sqrt{A_f M_2}$$

Рассмотрим пример. Пусть даны четырехзначные таблицы функции $y = \sin x$ (четырехзначные означает, что все четыре знака в таблице верные и что неустранимая погрешность $A_f = 0,5 \cdot 10^{-4}$). Требуется вычислить в одном из узлов функцию $y = \cos x$ по формуле численного дифференцирования на левый край и оценить количество верных знаков в результате. Каким следует взять оптимальный шаг $h = h_{опт}$ в формуле численного дифференцирования? Табличный шаг аргумента x предполагается равным 0,001.

В задаче, прежде всего, нужно определить при данной информации $h_{опт}$. По выведенной выше формуле, взяв в качестве $M_2 = 1$, получаем $h_{опт} \approx 0,014$, т.е. 14 табличных шагов. При этом $A_{пол}(h_{опт}) = 0,014$, т.е. всего один верный знак после запятой. И этот результат при данной информации улучшить за счет выбора шага нельзя.

8. ЧИСЛЕННОЕ ИНТЕГРИРОВАНИЕ

В отличие от численного дифференцирования, численное интегрирование – корректная задача. Эта одна из самых старых практических задач математики, её история отразилась в терминологии (квадратура).

8.1. Интерполяционные квадратурные формулы

Основная задача – посчитать приближенно величину определенного интеграла, используя вычисление значений подинтегральной функции.

Формула вида

$$\int_a^b f(x)dx \approx \sum_{k=0}^{n} A_k f(x_k)$$

называется *квадратурной*.

Один из основных подходов состоит в замене подинтегральной функции интерполяционным многочленом и его последующем интегрировании.

Пусть $f(x) \approx L_n(x)$, запишем интерполяционный многочлен в форме Лагранжа

$$L_n(x) = \sum_{k=0}^{n} f(x_k) \prod_{j=0, j\neq k}^{n} \frac{x - x_j}{x_k - x_j},$$

подставим его в интеграл и поменяем местами интеграл и сумму. Получим

$$\int_a^b f(x)dx \approx \sum_{k=0}^{n} f(x_k) \int_a^b \prod_{j=0, j\neq k}^{n} \frac{x - x_j}{x_k - x_j} dx$$

Обозначим

$$A_k = \int_a^b \prod_{j=0, j\neq k}^{n} \frac{x - x_j}{x_k - x_j} dx$$

и получим квадратурную формулу, которая называется *интерполяционной квадратурной*.

Интерполяционные квадратуры обладают свойством, которое выделяет их из всех квадратур.

Теорема 13. *Для того, чтобы квадратурная формула была интерполяционной, необходимо и достаточно, чтобы она была точна для любого многочлена степени n и ниже.*

Необходимость. Пусть квадратура интерполяционная, т.е. получилась заменой подинтегральной функции её интерполяционным многочленом $L_n(x)$. Но если подинтегральная функция является многочленом степени n и ниже, то она совпадает с интерполяционным многочленом и квадратура становится точной.

Достаточность. Пусть квадратура точна для любого многочлена степени n и ниже. Возьмем в качестве такого многочлена

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

и получим формулу

$$\int_a^b Q_i(x)dx = \sum_{k=0}^{n} A_k Q_i(x_k)$$

Но $Q_i(x_k) = 0, k \neq i$, $Q_i(x_i) = 0$, и последняя сумма вырождается в одно слагаемое A_i. Подставляя в эту формулу $Q_i(x)$, получаем

$$\int_a^b \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j} dx = A_i,$$

что совпадает с определением интерполяционной квадратуры.

Замечание.

Введем важное определение алгебраической степени точности формулы. Число N называется *алгебраической степенью точности* формулы,

если 1) она точна для всех многочленов степени N и ниже, 2) среди многочленов степени $N+1$ найдется хотя бы один, для которого формула неточна.

С учетом этого определения и обозначения N теорему можно переформулировать так.

Для того, чтобы квадратурная формула была интерполяционной, необходимо и достаточно, чтобы $N \geq n$.

8.2. Погрешность интерполяционных квадратурных формул

Для краткости будем применять следующие обозначения

$$I[f] = \int_a^b f(x)dx, \;\; S[f] = \sum_{k=0}^{n} A_k f(x_k),$$

тогда по определению погрешность интегрирования

$$R_{инт}[f] = I[f] - S[f],$$

а для интерполяционных квадратур, у которых $S[f]$ получено интегрированием интерполяционного многочлена

$$R_{инт}[f] = I[f] - I[L_n] = I[f - L_n] = I[R_n].$$

Подставляя формулу для погрешности интерполяции получаем

$$R_{инт}[f] = \int_a^b \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_n(x)dx$$

Заметим, что ξ зависит от x, поэтому оценить интеграл в общем случае затруднительно. Однако, если мы перейдем к модулю погрешности чис-

ленного интегрирования, то получим оценку

$$|R_{инт}[f]| \leq \frac{M_{n+1}}{(n+1)!} \int_a^b |\omega_n(x)|dx$$

8.3. Элементарные квадратурные формулы (Формулы Ньютона-Котеса)

Элементарные квадратурные формулы (Формулы Ньютона-Котеса) – это широко распространенные интерполяционные квадратуры при небольшом количестве узлов и при их стандартном расположении.

Их изучение начнем с $n = 0$ (один узел).

Формулы прямоугольников. Рассмотрим три случая.

А. $x_0 = a$. Тогда $L_0(x) = f(x_0) = f(a)$,

$$\int_a^b f(a)dx = f(a)(b-a)$$

и формула имеет вид

$$\int_a^b f(x)dx \approx f(a)(b-a),$$

эта формула называется *элементарной формулой левых прямоугольников*.

Алгебраическая степень точности $N = 0$, как можно проверить по определению.

Замечание. Геометрическая интерпретация этого и других рассмотренных ниже методов обычно рассматривается на практике, поэтому рисунки с иллюстрациями методов приводятся в методической разработке по практическим и лабораторным работам дисциплины "Численные методы".

Б. $x_0 = b$. Тогда $L_0(x) = f(b)$,

$$\int_a^b f(b)dx = f(b)(b-a)$$

и формула имеет вид

$$\int_a^b f(x)dx \approx f(b)(b-a),$$

эта формула называется *элементарной формулой правых прямоугольников*.

Алгебраическая степень точности $N = 0$.

В. $x_0 = \frac{a+b}{2}$. Тогда $L_0(x) = f(\frac{a+b}{2})$,

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

и формула имеет вид

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

эта формула называется *элементарной формулой средних прямоугольников*.

Алгебраическая степень точности $N = 1$. Это можно проверить, подставляя в формулу последовательно 1, x, x^2.

Формула трапеций. Пусть теперь $n = 1$ (два узла), $x_0 = a$, $x_1 = b$. Тогда $L_1(x) = f(a) + f(a,b)(x-a) = f(a) + \frac{f(b)-f(a)}{b-a}(x-a)$,

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

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

и формула имеет вид

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

Эта формула называется *элементарной формулой трапеций.*

Алгебраическая степень точности $N = 1$.

Формула Симпсона. Пусть $n = 2$ (три узла), $x_0 = a$, $x_1 = \frac{a+b}{2}$, $x_2 = b$. Тогда $L_2(x) = f(a) + f(a, \frac{a+b}{2})(x-a) + f(a, \frac{a+b}{2}, b)(x-a)(x-\frac{a+b}{2})$.

Чтобы не проделывать дальнейшее громоздкое расписывание разделенных разностей и последующее интегрирование, воспользуемся методом неопределенных коэффициентов. Запишем формулу

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

Подставим $f(x) \equiv 1$, получим равенство

$$b - a = A + B + C$$

Подставим $f(x) = x - a$, получим равенство

$$\frac{(b-a)^2}{2} = B\frac{(b-a)}{2} + C(b-a)$$

Подставим $f(x) = (x-a)^2$, получим равенство

$$\frac{(b-a)^3}{3} = B\frac{(b-a)^2}{4} + C(b-a)^2$$

Решая эту систему, получим $A = \frac{1}{6}(b-a)$, $B = \frac{4}{6}(b-a)$, $C = \frac{1}{6}(b-a)$, т.е. формула приобретает вид

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

эта формула называется элементарной интерполяционной формулой парабол или *элементарной формулой Симпсона.*

Алгебраическая степень точности $N = 3$. Это можно проверить, подставляя в формулу последовательно 1, x, x^2, x^3, x^4.

Формула "3/8". Пусть $n = 3$, четыре равноотстоящих узла, расстояние между которыми $\frac{b-a}{3}$, тогда $x_0 = a$, $x_1 = \frac{2a+b}{3}$, $x_2 = \frac{a+2b}{3}$, $x_3 = b$. Запишем формулу с неопределёнными коэффициентами

$$\int_a^b f(x)dx \approx Af(a) + Bf(\frac{2a+b}{3}) + Cf(\frac{a+2b}{3}) + Df(b).$$

Коэффициенты определяются, подставляя многочлены низших степеней, в результате $A = D = \frac{b-a}{8}$, $B = C = \frac{3(b-a}{)}8$.

Полученная формула

$$\int_a^b f(x)dx \approx \frac{b-a}{8}(f(a) + 3f(\frac{2a+b}{3}) + 3f(\frac{a+2b}{3}) + f(b))$$

называется *формулой "3/8".*

Заметим, что с увеличением числа узлов элементарные интерполяционные квадратурные формулы всё усложняются, что создает препятствия для их практического применения. Для повышения точности получил распространение другой подход – составные квадратурные формулы.

8.4. Погрешность элементарных интерполяционных квадратурных формул

Рассмотрим **погрешность элементарной формулы левых прямоугольников**. Как уже отмечалось, для интерполяционных квадратурных формул

$$R_{инт}[f] = I[R_n] = \int_a^b R_n(x)dx,$$

в частности, для элементарной формулы левых прямоугольников

$$R_{инт}[f] = I[R_n] = \int_a^b R_0(x)dx = \int_a^b f(x,a)(x-a)dx.$$

Воспользуемся известной из математического анализа теоремой о среднем, которая утверждает, что

$$\int_a^b g(x)h(x)dx = g(c)\int_a^b h(x)dx,\ c \in [a,b],$$

если $g(x)$ – непрерывная функция, $h(x)$ – непрерывная функция, сохраняющая знак на отрезке $[a,b]$.

Тогда, т.к. $(x-a) \geq 0$ на отрезке $[a,b]$, то по теореме о среднем

$$\int_a^b f(x,a)(x-a)dx = f(c,a)\int_a^b (x-a)dx = f(c,a)\frac{(b-a)^2}{2},\ c \in [a,b],$$

а используя связь разделенной разности и производной, получаем формулу для погрешности

$$R_{инт}[f] = f'(\eta)\frac{(b-a)^2}{2},\ \eta \in [a,b].$$

и её оценку

$$|R_{инт}[f]| \leq \frac{M_1}{2}(b-a)^2.$$

Аналогичным образом выводится **погрешность элементарной формулы правых прямоугольников**

$$R_{инт}[f] = -f'(\eta)\frac{(b-a)^2}{2},\ \eta \in [a,b].$$

и её оценка

$$|R_{инт}[f]| \leq \frac{M_1}{2}(b-a)^2.$$

Рассмотрим **погрешность элементарной формулы трапеций**.

$$R_{инт}[f] = I[R_n] = \int_a^b R_1(x)dx = \int_a^b f(x,a,b)(x-a)(x-b)dx.$$

Используя теорему о среднем (т.к. $(x-a)(x-b) \leq 0$ на отрезке $[a,b]$) и считая интеграл, получаем

$$\int_a^b f(x,a,b)(x-a)(x-b)dx = f(c,a,b)\int_a^b (x-a)(x-b)dx =$$

$$= -f(c,a,b)\frac{(b-a)^3}{6},\ c \in [a,b]$$

Используя связь разделенной разности второго порядка и соответствующей производной, получаем формулу для погрешности

$$R_{инт}[f] = -f''(\eta)\frac{(b-a)^3}{12},\ \eta \in [a,b].$$

и её оценку

$$|R_{инт}[f]| \leq \frac{M_2}{12}(b-a)^3.$$

Рассмотрим **погрешность элементарной формулы средних прямоугольников**. Прием, основанный на прямом интегрировании погрешности интерполяции не проходит, т.к. теорема о среднем неприменима, ибо $x - \frac{a+b}{2}$ не сохраняет знак на отрезке. Оценка модуля погрешности на этом пути может быть получена, но она слишком завышена. Применим прием, основанный на повышенной алгебраической степени точности формулы $N = 1$.

Пусть $H_1(x)$ – интерполяционный многочлен Эрмита, построенный по одному двухкратному узлу $x_0 = \frac{a+b}{2}$, т.е. построенный по интерполяционным данным $f(\frac{a+b}{2})$, $f'(\frac{a+b}{2})$. Так как в узле $x_0 = \frac{a+b}{2}$, функции $f(x)$ и $H_1(x)$ совпадают, то для формулы средних прямоугольников $S[f] = S[H_1]$.

Так как $H_1(x)$ – многочлен первой степени, а алгебраическая степень точности формулы $N = 1$, то $I[H_1] = S[H_1]$.

Получаем цепочку соотношений

$$R_{инт}[f] = I[f] - S[f] = I[f] - S[H_1] = I[f] - I[H_1] =$$

$$= I[f - H_1] = I[R_1].$$

Посчитаем интеграл от погрешности интерполяции многочленом $H_1(x)$, используя теорему о среднем

$$R_{инт}[f] = I[R_1] = \int_a^b f(x, \frac{a+b}{2}, \frac{a+b}{2})(x - \frac{a+b}{2})^2 dx =$$

$$= f(c, \frac{a+b}{2}, \frac{a+b}{2}) \int_a^b (x - \frac{a+b}{2})^2 dx = \frac{f''(\eta)}{2} \frac{(b-a)^3}{12}, \ \eta \in [a, b]$$

Таким образом получаем формулу для погрешности

$$R_{инт}[f] = f''(\eta) \frac{(b-a)^3}{24}, \ \eta \in [a, b].$$

и её оценку

$$|R_{инт}[f]| \leq \frac{M_2}{24}(b-a)^3.$$

Аналогичный прием применим для вывода **погрешности элементарной формулы Симпсона**, которая обладает повышенной алгебраической степени точности формулы $N = 3$.

Пусть $H_3(x)$ – интерполяционный многочлен Эрмита, построенный по двухкратному узлу $x_1 = \frac{a+b}{2}$, и однократным узлам $x_0 = a$ и $x_2 = b$ т.е. построенный по интерполяционным данным $f(a), f(\frac{a+b}{2}), f'(\frac{a+b}{2}), f(b)$. Так как в узлах $f(x)$ и $H_3(x)$ совпадают, то для формулы Симпсона $S[f] = S[H_3]$. Так как $H_3(x)$ – многочлен третьей степени, а алгебраическая степень точности формулы $N = 3$, то $I[H_3] = S[H_3]$.

Получаем цепочку соотношений

$$R_{инт}[f] = I[f] - S[f] = I[f] - S[H_3] = I[f] - I[H_3] =$$

$$= I[f - H_3] = I[R_3].$$

Посчитаем интеграл от погрешности интерполяции многочленом $H_1(x)$, используя теорему о среднем

$$R_{инт}[f] = \int_a^b f(x, a, \frac{a+b}{2}, \frac{a+b}{2}, b)(x-a)(x-b)(x-\frac{a+b}{2})^2 dx =$$

$$= f(c, a, \frac{a+b}{2}, \frac{a+b}{2}, b) \int_a^b (x-a)(x-b)(x-\frac{a+b}{2})^2 dx =$$

$$= -\frac{f''''(\eta)}{24}\frac{(b-a)^5}{120},\ \eta \in [a,b]$$

Последний интеграл лучше считать по частям.

Таким образом получаем формулу для погрешности

$$R_{инт}[f] = -f''''(\eta)\frac{(b-a)^5}{2880},\ \eta \in [a,b].$$

и её оценку

$$|R_{инт}[f]| \leq \frac{M_4}{2880}(b-a)^5.$$

8.5. Составные квадратурные формулы

Составные квадратурные формулы получаются путем разбиения отрезка интегрирования на равные части и применения на каждом отрезке одинаковых элементарных квадратурных формул. Рассмотрим этот процесс на примере **составной формулы левых прямоугольников**.

Пусть требуется посчитать

$$I[f] = \int_a^b f(x)dx.$$

Разобьем отрезок интегрирования на m частей с шагом $h = \frac{b-a}{m}$, введем точки $x_i = a+ih,\ i = 0, 1, \ldots, m$ разобьем интеграл на сумму m интегралов и для каждого применим элементарную формулу левых прямоугольников:

$$\int_a^b f(x)dx = \sum_{i=0}^{m-1} \int_{x_i}^{x_{i+1}} f(x)dx \approx \sum_{i=0}^{m-1} \int_{x_i}^{x_{i+1}} f(x_i)dx = h\sum_{i=0}^{m-1} f(x_i)$$

т.е.

$$I[f] \approx h\sum_{i=0}^{m-1} f(x_i).$$

Аналогичным образом выводится **составная формула правых прямоугольников**

$$I[f] \approx h\sum_{i=1}^{m} f(x_i),$$

составная формула средних прямоугольников

$$I[f] \approx h\sum_{i=0}^{m} f(x_i + \frac{h}{2}).$$

и **составная формула трапеций**

$$I[f] \approx h(\frac{f(a)+f(b)}{2} + \sum_{i=1}^{m-1} f(x_i)).$$

Выведем **составную формулу Симпсона**

$$\sum_{i=0}^{m-1} \int_{x_i}^{x_{i+1}} f(x)dx \approx \frac{h}{6}\sum_{i=0}^{m-1}(f(x_i) + 4f(x_i + \frac{h}{2}) + f(x_{i+1}))$$

Приводя подобные, получаем

$$I[f] \approx \frac{h}{6}(f(a) + f(b) + 2\sum_{i=1}^{m-1} f(x_i) + 4\sum_{i=0}^{m-1} f(x_i + \frac{h}{2}))$$

8.6. Погрешность составных квадратурных формул

Для того, чтобы получить погрешность составной квадратурной формулы, нужно просуммировать погрешности элементарных квадратурных формул на каждом отрезке. Так для **составной формулы левых прямоугольников**

$$R_{инт}[f] = \sum_{i=0}^{m-1} R^i_{инт}[f] = \sum_{i=0}^{m-1} f'(\eta_i)\frac{h^2}{2},\ \eta_i \in [x_i, x_{i+1}].$$

Отсюда, используя соотношение $mh = b - a$ получаем оценку погрешности

$$|R_{инт}[f]| \le \frac{M_1(b-a)}{2}h,\ M_1 = \max_{x\in[a,b]} |f'(x)|.$$

Эта оценка показывает, что *метод сходится*, т.е. погрешность стремится к нулю при стремлении шага h к нулю, если подинтегральная функция непрерывно дифференцируема.

Та же оценка погрешности справедлива для **составной формулы правых прямоугольников**.

Для **составной формулы средних прямоугольников** справедлива оценка погрешности

$$|R_{инт}[f]| \le \frac{M_2(b-a)}{24}h^2,\ M_2 = \max_{x\in[a,b]} |f''(x)|,$$

а для **составной формулы трапеций**

$$|R_{инт}[f]| \le \frac{M_2(b-a)}{12}h^2,\ M_2 = \max_{x\in[a,b]} |f''(x)|,$$

эти две формулы имеют второй порядок сходимости.

Составная формула Симпсона имеет оценку погрешности

$$|R_{инт}[f]| \le \frac{M_4(b-a)}{2880}h^4,\ M_4 = \max_{x\in[a,b]} |f''''(x)|.$$

8.7. Метод Рунге практической оценки погрешности

Выведенные оценки погрешностей на практике применять затруднительно, т.к. необходимо делать оценки соответствующих производных. Изложим другой метод оценки погрешностей, который позволяет по двум вычисленным значениям искомой величины (в данном случае интеграла) оценивать погрешность и уточнять результат.

Пусть требуется вычислить некоторую идеальную величину I, которая приближенно считается с помощью величины S_h, зависящей от параметра h. Обозначим погрешность через R_h, таким образом

$$I = S_h + R_h$$

Предположим, что погрешность представляется в виде асимптотического разложения

$$R_h = Ch^p + o(h^p),$$

где известен порядок p. Вычислим S_h при двух значения параметра h, например, S_h и $S_{h/2}$.

Из системы уравнений

$$\begin{cases} I = S_h + R_h \\ I = S_{h/2} + R_{h/2} \end{cases}$$

исключим неизвестную величину I, получим связь

$$S_{h/2} - S_h + R_{h/2} - R_h = 0.$$

С другой стороны, отбрасывая слагаемые более высокого порядка чем h^p, имеем представление

$$R_h \approx Ch^p,\ R_{h/2} \approx C(h/2)^p,$$

т.е.

$$R_h \approx 2^p R_{h/2}.$$

Отсюда получаем

$$R_{h/2} \approx \frac{S_{h/2} - S_h}{2^p - 1},$$

$$R_h \approx \frac{2^p}{2^p - 1}(S_{h/2} - S_h),$$

эти формулы называются *формулами Рунге практической оценки погрешности.* Они не только позволяют оценивать погрешность, но и уточнять результат:

$$I \approx S_{h/2} + \frac{S_{h/2} - S_h}{2^p - 1}$$

Рассмотрим применение этого метода к одной из составных формул, скажем к составной формуле левых прямоугольников. Сначала нужно доказать, что справедливо асимптотическое разложение погрешности.

Имеем

$$R_h = R_{инт}[f] = \sum_{i=0}^{m-1} f'(\eta_i)\frac{h^2}{2} = \frac{h}{2}(\sum_{i=0}^{m-1} f'(\eta_i)h) =$$

$$= \frac{h}{2}(\int_a^b f'(x)dx + O(h)),$$

т.е. для составной формуле левых прямоугольников имеет место асимптотическое разложение с $p = 1$. Формулы Рунге для составной формуле левых прямоугольников, также как и для составной формуле правых прямоугольников

$$R_{h/2} \approx S_{h/2} - S_h,$$

$$R_h \approx 2(S_{h/2} - S_h).$$

Для составной формулы трапеций и составной формулы средних прямоугольников, где $p = 2$ формулы Рунге имеют вид

$$R_{h/2} \approx \frac{(S_{h/2} - S_h)}{3},$$

$$R_h \approx \frac{4(S_{h/2} - S_h)}{3}.$$

Для составной формулы Симпсона, где $p = 4$ формулы Рунге имеют вид

$$R_{h/2} \approx \frac{(S_{h/2} - S_h)}{15},$$
$$R_h \approx \frac{16(S_{h/2} - S_h)}{15}.$$

8.8. Формулы наивысшей алгебраической степенью точности

В дальнейшем будем рассматривать задачу о вычислении интеграла с весом

$$I[f] = \int_a^b p(x)f(x)dx,$$

с помощью квадратурных формул

$$\int_a^b p(x)f(x)dx \approx \sum_{k=0}^{n} A_k f(x_k)$$

Весовая функция, которая входит в коэффициенты, используется, например, при вычислении интегралов с особенностями, см. ниже.

Интерполяционные квадратурные формулы обладают алгебраической степенью точности $N \geq n$, так как их коэффициенты A_k выбираются особым образом, раздел . Но можно выбирать и узлы интерполяции x_k, что еще может повысить алгебраическую степень точности.

Пример. Рассмотрим формулу

$$\int_{-1}^{1} f(x)dx \approx f(-\frac{\sqrt{3}}{3}) + f(\frac{\sqrt{3}}{3})$$

Её алгебраическая степень точности (как можно проверить) $N = 3$, а у аналогичной (с тем же количеством узлов) формулы трапеций

$$\int_{-1}^{1} f(x)dx \approx f(-1) + f(1)$$

алгебраическая степень точности $N = 1$.

Перейдем к нахождению условий, обеспечивающих наивысшую алгебраическую точность. Сначала заметим, что общее число параметров (коэффициентов и узлов) $2n + 2$, они обеспечивают условия, а многочлен степени $2n+1$ содержит столько же коэффициентов. Следует ожидать, что $N = 2n+1$. Это в самом деле так при определённых условиях, но доказать этот факт непросто.

Справедливо утверждение.

Теорема 14. *Для того, чтобы квадратурная формула с весом была точна для любого многочлена степени* $2n+1$ *и ниже, необходимо и достаточно, чтобы*

а) коэффициенты находились по формулам

$$A_k = \int_a^b p(x) \prod_{j=0, j\neq k}^{n} \frac{x - x_j}{x_k - x_j} dx$$

б) узлы такие, что многочлен $\omega_n(x) = (x - x_0)(x - x_1) \cdots (x - x_n)$ *ортогонален с весом любому многочлену* $g(x)$ *степени* n *и ниже. Ортогональность с весом многочленов* $\omega_n(x)$ *и* $g(x)$ *означает, что*

$$\int_a^b p(x)\omega_n(x)g(x)dx = 0.$$

Необходимость. Пусть квадратурная формула с весом точна для любого многочлена степени $2n + 1$ и ниже, возьмем в качестве такого много-

члена

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

и получим формулу

$$\int_a^b p(x) Q_i(x) dx = \sum_{k=0}^{n} A_k Q_i(x_k)$$

Но $Q_i(x_k) = 0, k \neq i$, $Q_i(x_i) = 0$, и последняя сумма вырождается в одно слагаемое A_i. Подставляя в эту формулу $Q_i(x)$, получаем

$$\int_a^b p(x) \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j} dx = A_i.$$

Возьмем любой многочлен $g(x)$ степени не выше n, тогда произведение $\omega_n(x)g(x)$ – многочлен, степени не выше $2n + 1$, следовательно для этого произведения квадратура с весом точна и, учитывая что $\omega_n(x_k) = 0$, получаем

$$\int_a^b p(x)\omega_n(x)g(x)dx = \sum_{k=0}^{n} A_k \omega_n(x_k) g(x_k) = 0$$

Достаточность. Сначала заметим, что т.к. коэффициенты A_k выбираются из условия замены интегрируемой функции её интерполяционным многочленом, то квадратура с весом точна для любого многочлена, степени n и ниже.

Возьмем любой многочлен $Q(x)$, степени не выше $2n+1$, поделим его на $\omega_n(x)$, т.е. представим в виде

$$Q(x) = g(x)\omega_n(x) + r(x),$$

где многочлены $g(x)$ и $r(x)$ имеют степени не выше n. Тогда, в силу условий а) и б)

$$\int_a^b p(x)Q(x)dx = \int_a^b p(x)g(x)\omega_n(x)dx + \int_a^b p(x)r(x)dx =$$

$$= \sum_{k=0}^{n} A_k r(x_k) = \sum_{k=0}^{n} A_k Q(x_k)$$

Доказанная теорема дает условия для формул наивысшей алгебраической степени точности (*квадратуры Гаусса*), но не решает вопрос о её существовании и единственности.

8.9. Существование и единственность квадратуры Гаусса

Представим многочлен $\omega_n(x)$ в виде разложения по степеням x:

$$\omega_n(x) = x^{n+1} + a_1 x^n + \ldots + a_n x + a_{n+1} = \Omega(x)$$

Обозначения $\Omega(x)$ будем применять, т.к. класс таких произвольных многочленов степени $n+1$ шире, чем ранее рассмотренные многочлены $\omega_n(x)$.

Распишем условие ортогональности с весом многочлена $\Omega(x)$ степеням x^l, $l = 0, 1, \ldots, n$:

$$\int_a^b p(x)(x^{n+1} + a_1 x^n + \ldots + a_n x + a_{n+1}) x^l dx = 0,\ l = 0, 1, \ldots, n$$

Получили линейную неоднородную систему относительно коэффициентов a_i. Эта система имеет единственное решение тогда и только тогда, когда соответствующая однородная система

$$\int_a^b p(x)(a_1 x^n + \ldots + a_n x + a_{n+1}) x^l dx = 0,\ l = 0, 1, \ldots, n$$

имеет только тривиальное решение. Умножим l-ое уравнение последней системы на a_l и сложим все уравнения, получим одно

$$\int_a^b p(x)(a_1 x^n + \ldots + a_n x + a_{n+1})^2 dx = 0.$$

Установим условия, при котором это уравнение имеет лишь тривиальное решение. Для этого рассмотрим вспомогательное утверждение.

Лемма. Пусть $p(x)$ непрерывная неотрицательная функция, такая, что $\int\limits_a^b p(x)dx > 0$, а многочлен $Q(x)$ неотрицателен на $[a,b]$. Тогда из условия $\int\limits_a^b p(x)Q(x)dx = 0$ следует $Q(x) \equiv 0$.

Доказательство. Так как $\int\limits_a^b p(x)dx > 0$, а функция $p(x)$ непрерывная неотрицательная, то найдется отрезок $[a_1,b_1] \subset [a,b]$, такой, что $p(x) > 0$ для всех $x \in [a_1,b_1]$.

Так как $Q(x)$ многочлен, то он имеет на отрезке $[a_1,b_1]$ конечное число корней $x_i, i = 1,\ldots,m$. Окружим каждый корень интервалом $(x-\varepsilon, x+\varepsilon)$ ($\varepsilon > 0$ мало) и введем множество $X_\varepsilon = [a_1,b_1] \setminus \bigcup\limits_{i=1}^{m}(x-\varepsilon, x+\varepsilon)$

На множестве ограниченном замкнутом множестве X_ε положительная непрерывная функция $Q(x)$ ограничена снизу: найдется $\delta > 0$, такое, что $Q(x) \geq \delta$ для всех $x \in [a_1,b_1]$.

Из равенства

$$0 = \int\limits_a^b p(x)Q(x)dx = \int\limits_{X_\varepsilon} p(x)Q(x)dx + \int\limits_{[a,b]\setminus X_\varepsilon} p(x)Q(x)dx$$

и неотрицательности подинтегральных функций следует, что

$$\int\limits_{X_\varepsilon} p(x)Q(x)dx = 0$$

Но тогда

$$0 = \int\limits_{X_\varepsilon} p(x)Q(x)dx \geq \delta \int\limits_{X_\varepsilon} p(x)dx > 0,$$

получили противоречие, следовательно $Q(x)$ имеет бесконечное число корней, т.е. $Q(x)$ – нулевой многочлен.

Если применить результат этой леммы к рассматриваемой задаче, то получаем утверждение

Теорема 15. *Пусть вес $p(x)$ непрерывная неотрицательная функция, такая, что $\int\limits_a^b p(x)dx > 0$, тогда существует единственный многочлен $\Omega(x) = x^{n+1} + a_1x^n + \ldots + a_nx + a_{n+1}$, ортогональный с весом степеням x^l, $l = 0, 1, \ldots, n$.*

Теорема 16. *В условиях предыдущей теоремы многочлен $\Omega(x)$ имеет ровно $n+1$ различных вещественных корней на отрезке $[a, b]$.*

Доказательство состоит из двух частей. Сначала докажем, что многочлен $\Omega(x)$ имеет $n+1$ корней на отрезке $[a, b]$. Предположим противное, т.е. что корней меньше, чем $n+1$, обозначим корни $x_0, x_1, \ldots, x_s$, $s \leq n-1$. Тогда

$$\Omega(x) = (x - x_0)(x - x_1) \cdots (x - x_s)r(x),$$

где $r(x)$ не обращается в ноль на отрезке $[a, b]$, для определённости будем считать, что $r(x) > 0$.

Обозначим $g(x) = (x-x_0)(x-x_1) \cdots (x-x_s)$, степень этого многочлена $s+1 \leq n$, воспользуемся ортогональностью с весом многочлену $\Omega(x)$, получим

$$\int\limits_a^b p(x)g^2(x)r(x)dx = 0,$$

применяя лемму, получим $r(x) \equiv 0$, противоречие, т.е. многочлен $\Omega(x)$ имеет $n+1$ корней на отрезке $[a, b]$.

Теперь докажем, что у многочлена $\Omega(x)$ кратных корней нет.

Предположим противное, т.е. что имеется корень, скажем x_0 кратности 2 (или выше). Тогда

$$\Omega(x) = (x - x_0)^2Q(x),$$

где степень многочлена $Q(x)$ равна $n-1$, воспользуемся ортогональностью

$Q(x)$ с весом многочлену $\Omega(x)$, получим

$$\int_a^b p(x)(x-x_0)^2 Q^2(x)dx = 0,$$

применяя лемму, получим $Q(x) \equiv 0$, противоречие, т.е. многочлен $\Omega(x)$ не имеет кратных корней.

Теоремы 15 и 16 дают условия существования и единственности квадратуры Гаусса.

8.10. Алгоритм построения квадратуры Гаусса

Подведем итог изложенному выше в виде алгоритма для построения квадратуры Гаусса.

Этап 1. Составляем систему

$$\int_a^b p(x)(x^{n+1} + a_1 x^n + \ldots + a_n x + a_{n+1})x^l dx = 0,\ l = 0, 1, \ldots, n$$

и вычисляем её коэффициенты.

Этап 2. Решаем эту систему, определяя $a_i, i = 0, 1, \ldots, n$.

Этап 3. Находим корни $x_k, i = 0, 1, \ldots, n$ многочлена $\omega_n(x) = x^{n+1} + a_1 x^n + \ldots + a_n x + a_{n+1}$.

Этап 4. Находим коэффициенты $A_k, i = 0, 1, \ldots, n$ квадратуры Гаусса

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

например, методом неопределенных коэффициентов.

Проиллюстрируем примером. Построим квадратуру Гаусса вида

$$\int_{-1}^{1} f(x)dx \approx A_0 f(x_0) + A_1 f(x_1).$$

Так как $n = 1$, составим многочлен $\omega_1(x) = x^2 + a_1 x + a_2$ и запишем систему из условия ортогональности этого многочлена 1 и x:

$$\begin{cases} \int\limits_{-1}^{1} (x^2 + a_1 x + a_2) dx = 0 \\ \int\limits_{-1}^{1} (x^2 + a_1 x + a_2) x dx = 0 \end{cases}$$

Считая интегралы, получаем

$$\begin{cases} 2a_2 dx = \frac{2}{3} \\ \frac{2}{3} a_1 = 0 \end{cases}$$

Находим корни уравнения $x^2 + \frac{1}{3} = 0$, получаем $x_0 = -\frac{\sqrt{3}}{3}$, $x_1 = \frac{\sqrt{3}}{3}$.

Записываем формулу

$$\int\limits_{-1}^{1} f(x) dx \approx A_0 f(-\frac{\sqrt{3}}{3}) + A_1 f(\frac{\sqrt{3}}{3})$$

и методом неопределенных коэффициентов находим $A_0 = A_A = 1$. Алгебраическая степень точности $N = 2n + 1 = 3$.

8.11. Погрешность квадратуры Гаусса

Будем, как и раньше, для краткости обозначать

$$I[f] = \int\limits_{a}^{b} p(x) f(x) dx, \; S[f] = \sum_{k=0}^{n} A_k f(x_k),$$

Пусть $H_{2n+1}(x)$ – интерполяционный многочлен Эрмита, построенный по двукратным узлам $x_0, x_1, \ldots, x_n$, т.е. построенный по интерполяционным данным $f(x_0)$,$f'(x_0)$, $\ldots$, $f(x_n)$, $f'(x_n)$. Так как в узлах функции $f(x)$ и $H_{2n+1}(x)$ совпадают, то для квадратуры Гаусса $S[f] = S[H_{2n+1}]$. Так как $H_{2n+1}(x)$ – многочлен степени $2n + 1$, а алгебраическая степень точности формулы $N \geq 2n + 1$, то $I[H_{2n+1}] = S[H_{2n+1}]$.

Получаем цепочку соотношений для погрешности интегрирования

$$R_{инт}[f] = I[f] - S[f] = I[f] - S[H_{2n+1}] = I[f] - I[H_{2n+1}] =$$

$$= I[f - H_{2n+1}] = I[R_{2n+1}].$$

Посчитаем интеграл от погрешности интерполяции многочленом $H_{2n+1}(x)$, используя теорему о среднем

$$R_{инт}[f] = \int_a^b p(x) f(x, x_0, x_0, \cdots, x_n, x_n)(x - x_0)^2 \cdots (x - x_n)^2 dx =$$

$$= f(c, x_0, x_0, \cdots, x_n, x_n) \int_a^b p(x)\omega_n^2(x)dx =$$

$$= \frac{f^{(2n+2)}(\eta)}{(2n+2)!} \int_a^b p(x)\omega_n^2(x)dx, \ \eta \in [a, b]$$

Таким образом, получаем формулу для погрешности

$$R_{инт}[f] = \frac{f^{(2n+2)}(\eta)}{(2n+2)!} \int_a^b p(x)\omega_n^2(x)dx, \ \eta \in [a, b]$$

и её оценку

$$|R_{инт}[f]| \leq \frac{M_{2n+2}}{(2n+2)!} \int_a^b p(x)\omega_n^2(x)dx.$$

Из этих оценок вытекают интересные следствия.

Замечание 1. Алгебраическая степень точности квадратуры Гаусса $N = 2n + 1$.

Нужно показать, что для многочленов степени $2n + 2$ формула не точна. Возьмем такой многочлен $Q(x) = x^{2n+2}$. Тогда его производная порядка $2n + 2$ равна $(2n + 2)!$, а $\int_a^b p(x)\omega_n^2(x)dx > 0$, т.е. $R_{инт}[Q] \neq 0$.

Замечание 2. Коэффициенты квадратуры Гаусса A_k положительны.

Для доказательства возьмем многочлен

$$Q_i(x) = [\prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j}]^2$$

степени $2n + 2$. Тогда

$$0 < \int_a^b p(x)Q_i(x)dx = \sum_{k=0}^{n} A_k Q_i(x_k) = A_i,$$

что и требовалось доказать.

8.12. Вычисление интегралов с особенностями

Под интегралами с особенностями понимаются интегралы от неограниченных функций или интегралы на бесконечных промежутках интегрирования. Кроме того, особенностью называется недостаточная гладкость подинтегральной функции при применении той или иной квадратурной формулы, так, например, для применения составной формулы Симпсона требуется непрерывность четвертой производной.

Рассмотрим на примерах некоторые приемы, позволяющие устранять особенности.

Аналитические способы.

1. Требуется вычислить

$$\int_0^{\text{П}/2} \ln \sin x dx,$$

имеющий особенность в нуле. Добавим и вычтем $\sin x$ к подинтегральной функции и получим

$$\int_0^{\text{П}/2} \ln \sin x dx = \int_0^{\text{П}/2} \ln x dx + \int_0^{\text{П}/2} \ln \frac{\sin x}{x} dx.$$

Первый интеграл может быть вычислен аналитически, а второй особенности не содержит и может быть вычислен численно.

2. Чтобы сдвинуть особенность в старшую производную, часто применяют формулу интегрирования по частям. Рассмотрим

$$\int_0^a x^\alpha g(x)dx,$$

где функция $g(x)$ "хорошая" (достаточное число раз дифференцируемая). При $-1 < \alpha < 0$ подинтегральная функция неограничена, при $0 < \alpha < 1$ подинтегральная функция ограничена, но её первая производная неограничена и так далее. Проинтегрируем по частям

$$\int_0^a x^\alpha g(x)dx = \frac{1}{\alpha+1}x^{\alpha+1}g(x)|_0^a - \frac{1}{\alpha+1}\int_0^a x^{\alpha+1}g'(x)dx,$$

таким образом сдвинули особенность на единицу в старшую производную.

Метод усечений.

Суть метода состоит в том, что мы отделяемся от особенности с нужной точностью. В качестве примера рассмотрим задачу о вычислении

$$\int_2^\infty e^{-x^2}dx,$$

с точностью $\varepsilon = 0,0001$. Имеем

$$\int_2^\infty e^{-x^2}dx = \int_2^B e^{-x^2}dx + \int_B^\infty e^{-x^2}dx,$$

Последний интеграл дает погрешность усечения.

Сделаем оценку

$$\int_B^\infty e^{-x^2}dx \le \int_B^\infty \frac{2x}{2A}e^{-x^2}dx = \frac{1}{2A}e^{-A^2} < \varepsilon/2.$$

Вычисления показывают, что при $B = 3$ эта оценка выполняется.

Оценка погрешности составной формулы Симпсона показывает, что если взять $n = 8$, то выполняется

$$\int_2^3 e^{-x^2} dx < \varepsilon/2.$$

Введение весовой функции. Мультипликативный метод выделения особенностей.

Если подинтегральная функция есть произведение двух функций, "хорошей" и содержащей особенность, то имеет смысл рассматривать формулы численного интегрирования с весом, вес включается в коэффициенты квадратуры. Например, для вычисления

$$\int_0^1 \sqrt{x} f(x) dx$$

можно использовать интерполяционные квадратурные формулы с весом

$$\int_0^1 \sqrt{x} f(x) dx \approx \sum_{k=0}^{n} A_k f(x_k),\ A_k = \int_0^1 \sqrt{x} \prod_{j=0, j\neq k}^{n} \frac{x - x_j}{x_k - x_j} dx$$

К этому же классу методов относятся специальные *формулы для быстро осцилирующих функций*. Требуется вычислить

$$\int_a^b g(x) \sin \omega x dx.$$

Попробуем применить какую-либо составную формулу, например, составную формулу средних прямоугольников

$$\int_a^b g(x) \sin \omega x dx \approx h \sum_{i=0}^{n-1} g(x_{i+1/2}) \sin \omega x_{i+1/2}$$

В оценку погрешности этой формулы входит производная, однако

$$f'(x) = (g(x)\sin\omega x)' = g'(x)\sin\omega x + \omega g(x)\cos\omega x$$

и эта функция при больших ω может принимать большие значения.

Выведем аналог составной формулы средних прямоугольников заменяя константой в средней точке не всю подинтегральную функцию, а только $g(x)$, включая $\sin\omega x$ в весовую функцию.

$$\int_a^b g(x)\sin\omega x dx = \sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} g(x)\sin\omega x dx \approx$$

$$\approx \sum_{i=0}^{n-1} g(x_{i+1/2})\int_{x_i}^{x_{i+1}} \sin\omega x dx =$$

$$= -\frac{1}{\omega}\sum_{i=0}^{n-1} g(x_{i+1/2})(\cos\omega x_{i+1} - \cos\omega x_i) =$$

$$= \frac{2}{\omega}\sin\omega\frac{h}{2}\sum_{i=0}^{n-1} g(x_{i+1/2})\sin\omega x_{i+1/2}$$

При этом погрешность оценивается величиной

$$|R_{инт}[f]| \le \sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} |g(x, x_{i+1/2})||x - x_{i+1/2}||\sin\omega x| dx \le$$

$$\le \frac{M_1[g](b-a)}{4}h,\ M_1[g] = \max_{x\in[a,b]}|g'(x)|$$

которая стремится к нулю при шаге h стремящемся к нулю равномерно по ω.

Аддитивный метод устранения особенностей.

Идея метода состоит в том, чтобы представить исходную подинтегральную функцию в виде суммы $f(x) = \varphi(x) + \psi(x)$, где интеграл от $\varphi(x)$, содержащей те же особенности, что и исходная функция, считается

аналитически, а функция $\psi(x)$ не содержит особенности и интеграл от неё считается численно. Основным инструментом при этом является тейлоровское разложение в окрестности особенности.

Рассмотрим пример. Требуется сдвинуть особенность в третью производную для подсчета

$$\int_0^1 \frac{\sin x}{x^{1/2}(1+x^2/2)}dx.$$

Разложим функции в нуле

$$\sin x = x - \frac{x^3}{6} + \cdots$$

$$\frac{1}{(1+x^2/2)} = 1 - \frac{x^2}{2} + \frac{x^4}{4} + \cdots$$

$$\frac{\sin x}{x^{1/2}(1+x^2/2)} = (x^{1/2} - \frac{x^{5/2}}{6} + \cdots)(1 - \frac{x^2}{2} + \frac{x^4}{4} + \cdots) =$$

$$= x^{1/2} - \frac{2}{3}x^{5/2} + \cdots$$

Обозначим

$$\varphi(x) = x^{1/2} - \frac{2}{3}x^{5/2}$$

$$\psi(x) = \frac{\sin x}{x^{1/2}(1+x^2/2)} - x^{1/2} + \frac{2}{3}x^{5/2}$$

Интеграл $\varphi(x)$ считается аналитически, а функция $\psi(x)$ по крайней мере дважды непрерывно дифференцируема и интеграл от неё может быть вычислен, например, с помощью составной формулы трапеций.

ЛИТЕРАТУРА

1. *Бабенко К.И.* Основы численного анализа. М.: Наука, 1986. 452 с.

2. *Бахвалов Н.С.* Численные методы. М.: Наука, 1973. 632 с.

3. *Бахвалов Н.С., Жидков Н.П., Кобельков Г.М.* Численные методы. М.: Наука, 1987. 598 с.

4. *Березин И.С., Жидков Н.П.* Методы вычислений. Т. 1. М.: Физматгиз, 1959. 464 с.

5. *Березин И.С., Жидков Н.П.* Методы вычислений. Т. 2. М.: Наука, 1966. 430 с.

6. *Вержбицкий В.М.* Численные методы. Линейная алгебра и нелинейные уравнения. М.: ОНИКС 21 век, 2005. 432 с.

7. *Вержбицкий В.М.* Численные методы. Математический анализ и обыкновенные дифференциальные уравнения. М.: ОНИКС 21 век, 2005. 400 с.

8. *Волков Е.А.* Численные методы. М: Наука, 1982. 248 с.

9. *Калиткин Н.Н.* Численные методы. 2-е издание. СПб.: БХВ-Петербург, 2011. 586 с.

10. *Крылов В.И., Бобков В.В., Монастырный П.И.* Начала теории вычислительных методов (в 5 томах). Интерполирование и интегрирование. Минск, Наука и техника, 1983. 287 с.

11. *Крылов В.И., Бобков В.В., Монастырный П.И.* Начала теории вычислительных методов (в 5 томах). Линейная алгебра и нелинейные уравнения. Минск, Наука и техника, 1985. 279 с.

12. *Марчук Г.И.* Методы вычислительной математики. М.: Наука, 1989. 608 с.

13. *Петров И.Б., Лобанов А.И.* Лекции по вычислительной математике. М.: БИНОМ, 2006. 524 с.

14. *Самарский А.А., Гулин А.В.* Численные методы. М.: Наука, 1989. 430 с.

15. *Фадеев А.К., Фадеева В.Н.* Вычислительные методы линейной алгебры. СПб.: Лань, 2002. 736 с.

16. *Формалев В.Ф., Ревизников Д.Л.* Численные методы. М.: ФИЗМАТЛИТ, 2006. 400 с.

Библиографический комментарий

Учебную литературу условно можно разбить на несколько групп. К первой группе относятся классические университетские учебники [4, 5, 2, 3, 14, 9, 10, 11], данный курс читается ближе всего к учебникам [14, 9]. Ряд книг [12, 1, 15] предназначен, прежде всего, для специалистов и они содержат много дополнительных сведений. Имеется очень много учебников для технических университетов, особенно появившихся в последнее время, отметим некоторые [6, 7, 13, 16], которые содержат описание разнообразных методов, но зачастую в них отсутствуют доказательства, например, совсем простой учебник [8].