ФИЗИЧЕСКИЙ
ФАКУЛЬТЕТ
МГУ ИМЕНИ
М.В. ЛОМОНОСОВА

teach-in
ЛЕКЦИИ УЧЕНЫХ МГУ

ЧИСЛЕННЫЕ МЕТОДЫ В ФИЗИКЕ

ПРИКЛОНСКИЙ ВЛАДИМИР ИВАНОВИЧ

ФИЗФАК МГУ

БЛАГОДАРИМ ЗА ПОДГОТОВКУ КОНСПЕКТА
СТУДЕНТКУ ФИЗИЧЕСКОГО ФАКУЛЬТЕТА МГУ
УЛЬЯНОВУ ЛИВИЮ-НИКОЛЬ

Содержание

1. Лекция 1. Вводная лекция **6**
1.1. Процесс математического моделирования. 6
1.2. Примеры постановки задачи вычисления 8
1.3. Погрешности задачи вычисления 10
1.4. Погрешности округления на t-разрядной ЭВМ. 12

2. Лекция 2. Интерполяция и приближение функций **14**
2.1. Постановка задачи интерполяции. Линейная интерполяция. 14
2.2. Полиномиальная интерполяция 15
2.3. Интерполяционный полином Лагранжа 16
2.4. Интерполяционный полином Ньютона 18
2.5. Погрешность полиномиальной интерполяции 20
2.6. Сходимость интерполяционного многочлена 22

3. Лекция 3. Сплайн-интерполяция **25**
3.1. Кубический сплайн 25

4. Лекция 4. Аппроксимация функций **30**
4.1. Существование и единственность наилучшего среднеквадратичного приближения 30
4.2. Метод наименьших квадратов(МНК) 33
4.3. Обработка экспериментальных кривых методом НК 33

5. Лекция 5. Вопросы численного дифференцирования и интегрирования функций. **37**
5.1. Квадратурные формулы Ньютона-Котесса 38
5.2. Частные случаи формул Ньютона-Котесса 39

6. Лекция 6.Квадратурные формулы Гаусса- Кристоффеля **44**
6.1. Выбор узлов квадратурной формулы 44
6.2. Веса квадратурной формулы 45
6.3. Простейший случай квадратурных формул Гаусса-Кристоффеля (формула средних прямоугольников) 45
6.4. Апостериорная оценка погрешности 47
6.5. Численное дифференцирование 48

7. Лекция 7.Решение нелинейных уравнений **50**
7.1. Метод деления отрезка пополам 50
7.2. Метод последовательного приближения. Теорема о непрерывном сжатии 51
7.3. Итерационные методы решения систем нелинейных уравнений 57

8. Лекция 8. Основные методы решения уравнений. Метода последовательного исключения Гаусса. **60**
8.1. Метод последовательного исключения Гаусса 63

8.2. LU - разложение невырожденной матрицы 65

9. Лекция 9. Итерационные методы решения систем линейных уравнений. Часть 1 68
9.1. LU-разложение ленточной матрицы 68
9.2. Итерационные методы решения СЛАУ 69
9.3. Основные итерационные методы 72

10. Лекция 10. Итерационные методы решения систем линейных уравнений. часть 2 74
10.1. Теорема Самарского 75
10.2. Достаточные условия сходимости простейших итерационных методов 76

11. Лекция 11. Алгебраическая проблема поиска собственных значений 79
11.1. Устойчивость невырожденной задачи нахождения собственных векторов и собственных значений 79
11.2. Метод парабол 81
11.3. Метод вращений (Якоби) 83
11.4. Оценка нормы матрицы 84

12. Лекция 12. Задачи минимизации 86
12.1. Минимизация функции одного переменного. Методы нулевого порядка. 86
12.2. Метод более высокого порядка 89
12.3. Минимизация функции многих переменных 90
12.4. Квадратичная функция аргумента $\vec{x}$ 91
12.5. Рельеф поверхности функции $\Psi(x)$. Линии уровня 92
12.6. Спуск по координатам. 93

13. Лекция 13. Методы минимизации 94
13.1. Метод покоординатного спуска. Продолжение 94
13.2. Метод покоординатного спуска в общем случае 95
13.3. Метод наискорейшего спуска 97
13.4. Методы второго порядка. Метод сопряженных градиентов 98

14. Лекция 14. Минимизация функционала 100
14.1. Сходимость последовательности значений функционала и последовательности аргументов 100
14.2. Задачи минимизации функционала 102
14.3. Метод Ритца 103

15. Лекция 15. Разностные методы решения задач математической физики. Часть 1 104
15.1. Аппроксимация разностной схемы 105
15.2. Устойчивость 105
15.3. Двуслойные разностные схемы 106
15.4. Сходимость разностной схемы 109
15.5. Задача построения сеточной аппроксимации 110

16. Лекция 16. Разностные методы решения задач математической физики. Часть 2 **113**
16.1. Одномерное уравнение теплопроводности 113
16.2. Устойчивость разностной схемы 115
16.3. Разностная схема крест 118

17. Лекция 17. Разностные методы решения задач математической физики. Часть 3 **120**
17.1. Многомерные разностные схемы для уравнения теплопроводности . . 121
17.2. Продольно-поперечная разностная схема 123
17.3. Устойчивость продольно-поперечной неявной схемы 124
17.4. Аппроксимация продольно-поперечной схемы 125

18. Лекция 18. Дополнительная **126**
18.1. Постановка задачи 126
18.2. Явная схема 126
18.3. Неявная схема 128

1. Лекция 1. Вводная лекция

1.1. Процесс математического моделирования.

Мы начинаем курс лекций по численным методам в физике. Этот курс традиционно читается весной для студентов 4 курса. Рассмотрим схематически этапы математического моделирования (рис. 1.1).

Цель математического моделирования состоит в том, чтобы построить на первом этапе "адекватную"математическую модель некоторого физического явления. При построении математической модели мы стараемся выделить наиболее важные и характерные зависимости, связи и ограничения, которые предъявляются к модели в процессе ее изучения. Как правило, для нас математическая модель выступает в виде системы нелинейных дифференциальных уравнений в частных производных (СНДУЧП). Соответствующая СНДУЧП описывает поведение интересующих нас функций, характеризующих физический процесс. Вдобавок к этому присутствуют соответствующие параметры, определяющие конкретные условия протекания процесса, а также характеристики, связанные с ограничениями, налагаемыми за счет наличия дополнительных границ и так далее. Мы стараемся отобразить многогранность физического процесса в математической модели. Дифференциальные уравнения в частных производных (ДУЧП) возникают из применения законов сохранения к рассматриваемому физическому явлению. Записывая дифференциальные формы законов сохранения в предельном переходе, мы приходим к ДУЧП, и в частном случае эти уравнения превращаются в обычные дифференциальные уравнения. Нелинейная зависимость вытекает из специфики протекающего процесса, так как далеко не всегда мы сталкиваемся с относительно простыми физическими явлениями, которые легко описываются линейными дифференциальными уравнениями.

Следующим этапом математического моделирования является проведение математического исследования модели. На этом шаге ставится вопрос о нахождении решения. При анализе модели на разных этапах может вкладываться различный смысл в понятие решение задачи. Традиционным примером является доказательство теоремы о существовании и единственности решения. Теорема в определенном смысле решает задачу, но не позволяет нам изучить качественное поведение решения и оценить его количественные характеристики. Относительно простые модели допускают возможность аналитического описания интересующего нас решения.

Часто при изучении модели мы можем применять приближенные методы решения такие как осреднение, изучение различных асимптотик и другими. Например, если протекающие процессы характеризуются временными шкалами мы можем расширить или наоборот сузить их. Тогда процессы, которые протекают слишком быстро при изменении временного масштаба, могут выйти на какое-то асимптотическое решение. И аналитические, и приближенные методы приводят нас к построению аналитического решения. Аналитические решения описывают достаточно широкий класс поведения соответствующей модели, в отличие от других методов, которые имеют дело с некоторым частным решением.

Наконец, для наиболее точных и сложных моделей основными методами решения являются численные методы решения с необходимостью проведения большого

объема вычислений на ЭВМ. Эти методы позволяют добиться хорошего количественного и даже качественного результата в описании модели. Правда, у них есть и принципиальные недостатки, как правило, речь идет о рассмотрении некоторого частного решения. С другой стороны, мы также понимаем, что и аналитическое решение впоследствии требует от нас получения конкретного решения с помощью численных методов. Для получения численного решения нам необходимо создать дискретную модель. Реализация дискретной модели дальше требует ее алгоритмизации на каком-то языке программирования.

Последним этапом моделирования является соотнесение полученных нами количественных результатов с результатами физического эксперимента. Сопоставление этих результатов позволяет нам сделать вывод о состоятельности модели. Хорошее согласование с экспериментом обычно свидетельствует о правильности выбора модели. В противном случае нужно более внимательно сосредоточиться на одном из пройденных этапов.

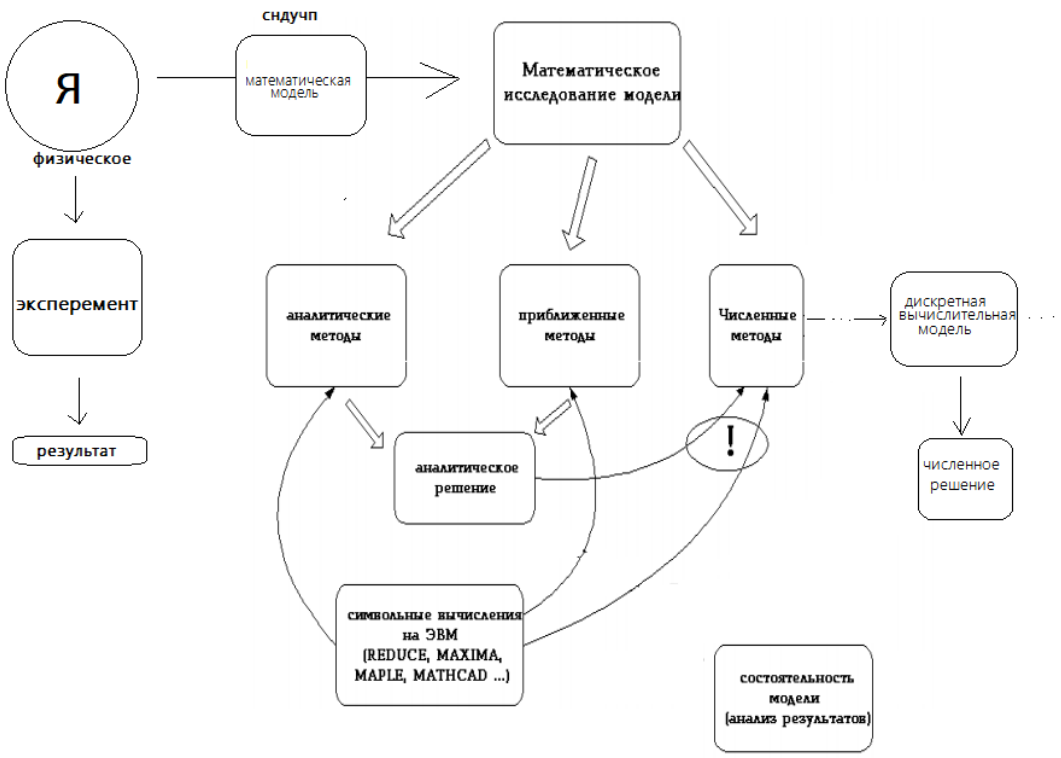

Рис. 1.1 – Схема этапов математического моделирования

Пусть мы хотим вычислить неизвестную величину y по известной величине x. Запишем самым общим образом зависимость интересующего нас результата от исходных данных как:

$$A = y(x) \tag{1}$$

где $y \in Y$, $x \in X$ — элементы соответствующих функциональных пространств; A — оператор, реализующий вычисления.

В первую очередь нас будут интересовать корректно поставленные задачи вычисления.

Задача вычисления (1) называется корректно поставленной по Адамару, если для любых входных данных из некоторого класса решение задачи существует, единственно и непрерывно зависит от входных данных задачи (или устойчиво по входным данным).

Поставленные к задаче требования являются вполне естественными. Для того чтобы численно решать задачу нужно быть уверенным, что ее решение существует. Рассматривая детерминированные задачи, мы понимаем что если сформулированы все условия определяющие протекания процесса, то повторение этих же условий должно дать нам тот же самый результат, отсюда вытекает требование единственности. Так как наши рассмотрения носят приближенный характер, как правило, входные данные неизвестны нам абсолютно точно, то естественно требование того, что малые возмущения входных данных должны приводить к малым возмущениям решения.

Сделаем несколько замечаний об устойчивости. Нас интересует решение задачи (1), соответствующее входным данным x. Реально мы имеем возмущенные входные данные с погрешностью δx, тем самым получаем возмущенное решение

$$y + \delta y = A(x + \delta x)$$

Эта погрешность входных данных порождает неустранимую погрешность решения

$$\delta y = A(x + \delta x) - A(x)$$

Устойчивость означает, что если мы контролируем погрешность входных данных, и эта величина стремится к нулю, то в устойчивых задачах вычисления соответствующая неустранимая погрешность также стремится к нулю.

$$\|\delta x\| \to 0 \Longrightarrow \|\delta y\| \to 0$$

Для решения задач мы пользуемся следующей оценкой

$$\|\delta y\| \leq C \cdot |\delta x\|$$

Часто с константой C связывают понятие обусловленности задачи. Если порядок близок к единице, то погрешности хорошо согласуются. Если константа C будет достаточно велика, то погрешность δy может оказаться неприемлемо большой, в этом случае говорят что задача плохо обусловлена.

Приведем несколько примеров постановки задачи вычисления.

1.2. Примеры постановки задачи вычисления

1. Задача нахождения корней полинома

Рассмотрим некоторый полином степени n в приведенном виде (старший коэффициент равен единице):

$$p_n(x) = x^n + a_1 x^{n-1} + a_2 x^{n-2} + ... + a_n, \quad \text{где } a_n, x \in \mathbb{C}$$

Так как мы рассматриваем многочлен в поле комплексных чисел, то по основной теореме алгебры мы можем гарантировать, что полином n-степени имеет ровно n корней. Требуется определить его корни. Пусть E^n - n—мерное комплексное евклидово пространство. Пусть компоненты некоторого вектора $\vec{z} = \{z_1, z_2, .., z_n\}$ являются корнями полинома $p_n(x)$, т.е

$$p_n(z_k) = 0, \quad k = \overline{1, n}$$

Тогда, опираясь на теорему Безу, мы можем написать разложение $p_n(x)$ в виде:

$$p_n(x) = (x - z_1)(x - z_2)...(x - z_n)$$

Отсюда мы получаем формулы Виета, отображающие связь между коэффициентами многочлена и его корнями.

$$a_k = (-1)^k \sigma_k$$

Здесь σ_k элементарные, симметричные относительно $z, z_1, z_2, .., z_n$ однородные функции k-го порядка

$$\begin{cases} \sigma_1 = z_1 + z_2 + ... + z_n \\ \sigma_2 = z_1 z_2 + z_1 z_3 + ... + z_{n-1} z_n \\ \dots \\ \sigma_n = z_1 z_2 ... z_n \end{cases}$$

Эти функции σ_k возникают при раскрытии скобок в разложении многочлена, и каждая из них содержит C_n^k слагаемых.

Таким образом, формулы Виета сопоставляют каждому вектору $z \in E^n$ вектор $\vec{a} = \{a_1, a_2, ..., a_n\}$ того же пространства, т.е. определяют отображение $V : E^n \Longrightarrow E^n$ пространства E^n на себя. Тем самым наша задача свелась к реализации соответствующего отображения V, то есть для заданного вектора $\vec{a}$ найти вектор $\vec{z} \in E^n$ такой, что

$$V(\vec{z}) = \vec{a} \tag{2}$$

Наша задача вычисления оказалась не в форме $y = A(x)$, нам необходимо построить обращения отображения 2

$$\vec{z} = V^{-1}(\vec{a})$$

Конструктивное построение V^{-1} составляет процесс реализации задачи вычисления по входным данным x построим y. Как мы видим, даже в этом конкретном случае соответствующая задача вычисления не задана нам явно.

В курсе высшей алгебры показано, что отображение V взаимно-однозначное и взаимно-непрерывное, т.е. задача (2) корректна.

2. Основная задача линейной алгебры

пусть дана матрица $A_(p \times q) = \|a_j^i\|_q^p$, тогда определено отображение

$$A : E^q \Longrightarrow E^P; \quad \vec{y} = A(\vec{x}), \quad \vec{y} \in E^p, \quad \vec{x} \in E^q$$

$\vec{x}, \vec{y}$ - столбцы соответствующих размерностей

Основная задача линейной алгебры состоит в том, чтобы по заданному вектору $\vec{f} \in E^p$ построить вектор $\vec{x} \in E^q$ такой, что

$$A\vec{x} = \vec{f} \tag{3}$$

Задача (3) представляет собой задачу решения системы линейных алгебраических уравнений — СЛАУ. Соответствующие теоремы, гарантирующие корректность поставленной задачи, сводятся к следующим случаям

1) Если матрица является квадратной ($p = q$), то из формул Крамера следует, что решение задачи (3) существует и единственно при любом f в случае, когда $det\, A \neq 0$.

2)В остальных случаях по теореме Кронекера-Капелли если ранг основной и расширенной матрицы системы (3) совпадает, то решение существует, но не единственно. В противном случае решение вовсе отсутствует, т.е. задача (3) в этих случаях некорректно поставлена.

3. Задача Коши для обыкновенного дифференциального уравнения

Требуется найти решение задачи Коши для обыкновенного дифференциального уравнения (ОДУ), отвечающее начальному условию $y(a) = c$

$$\begin{cases} \frac{dy}{dx} = f(x, y) \quad ay(a) = c \end{cases}$$

a, b - заданные числа; $f(x, y)$ - определена в полуполосе $\Pi = \{(x, y); \quad x \in [a, b]; \quad y \in (-\infty, +\infty)\}$

Обозначим множество всех решений задачи коши как K^o и определим на нем отображение

$$K(y(x)) = y(a) \quad \forall y \in K^o$$

Тогда, решение задачи Коши для ОДУ можно сформулировать так: по заданному числу c найти функцию $y(x)$ такую, что

$$K(y(x)) = c \tag{4}$$

В курсе дифференциальных уравнений доказана корректность задачи (4).

1.3. Погрешности задачи вычисления

Вернемся к задаче вычисления (1)

$$y = A(x)$$

Выделяют 4 основных источника погрешность результата вычисления

1. $\delta_1 y$ - Погрешность математической модели. Эта погрешность связана с допущениям при переходе от физического явления к его математической модели, она останется вне рамок нашего рассмотрения.

2. $\delta_2 y$ - Погрешность входных данных. Порождает неустранимую погрешность решения

$$\delta_2 y = A(x + \delta x) - A(x)$$

2. $\delta_3 y$ - Погрешность метода. Если задача (1) достаточно сложно численно реализуема, то можно решить более простую близкую к ней задачу

$$\overline{y} = \overline{A}(\overline{x}) \tag{5}$$

Нам необходимо перейти к новым функциональным пространствам $\overline{X}, \overline{Y}$ и суметь реализовать оператор $\overline{A}$. Величина

$$\delta_3 y = y - \overline{y} = A(x) - A(\overline{x})$$

и представляет собой погрешность метода. Естественно мы хотим, чтобы решение $\overline{y}$ было близко к решению y и соответственно погрешность метода стремилась бы к нулю.

2. $\delta_4 y$ - вычислительная погрешность(погрешность округления).

Пусть существует численная реализация $\overline{y}$, в процессе ее вычисления возникают дополнительные погрешности, как правило, связанные с округлением и мы получаем величину $\tilde{y}$.

$$\delta_4 y = \overline{y} - \tilde{y} = \overline{A}(\overline{x}) - \tilde{y}$$

Полезно сразу же сформулировать некоторые эмпирические правила, которых придерживаются при реализации задачи вычисления:

$$\|\delta_2 y\| = (2 \div 5)\|\delta_3 y\| \gg \|\delta_4 y\|$$

1) Погрешность метода $\delta_3 y$ должна быть меньше неустранимой погрешности $\delta_2 y$. То есть эти погрешности можно связать коэффициентом порядка единицы.

2) Вычислительная погрешность $\delta_4 y$ должна быть существенно меньше всех остальных погрешностей решения, т.е. расчет нужно вести с таким количеством значащих цифр, чтобы погрешность округления была существенно меньше всех остальных погрешностей.

Несколько слов о вычислительной погрешности. Мы будем считать, что результаты вычислений $\tilde{y}$, возмущенные вычислительными погрешностями, совпадают с реализацией точного алгоритма $\overline{A}$ на возмущенных входных данных

$$\tilde{y} = \overline{A}(\tilde{x})$$

Хоть это и является допущением, но оно характеризует достаточно широкий класс алгоритмов, предлагаемых для реализации задач вычисления. Тогда величина погрешности δ_4 сведется к устойчивости предложенного алгоритма

$$\delta_4 y = \overline{A}(\overline{x}) - \overline{A}(\tilde{x})$$

Если мы контролируем погрешность вносимую $\overline{x}$ и $\tilde{x}$, то устойчивость алгоритма $\overline{A}$ гарантирует нам малость величины $\delta_4 y$.

Теперь мы можем сформулировать наши основные задачи в рамках курса "Численных методов"

1) Конструирование дискретной модели $\overline{X}, \overline{Y}, \overline{A}$.

2) Разработка на ее основе соответствующих алгоритмов решения задачи вычисления (5).

3) Анализ погрешности метода $\delta_3 y$ и частично вычислительной погрешности $\delta_4 y$ алгоритма, реализующего вычисления $\overline{A}$.

1.4. Погрешности округления на t-разрядной ЭВМ.

В современных ЭВМ действительные числа представляются в т.н. форме с плавающей запятой, т.е. если само число a в позиционной системе счисления с основанием r записано в виде r-ичной дроби

$$a = sign\ a(a_n a_{n-1}...a_1 a_0 a_{-1} a_{-2}...)_r = \quad 0 \le a_i \le r-1$$

$$= sign\ a\left(a_n r^n + ... + a_1 r + a_0 + \frac{a_{-1}}{r} + \frac{a_{-2}}{r^2} + ...\right)_r$$

то такую форму записи числа f называют представлением с фиксированной запятой. Здесь a_k – r-ичные цифры. В привычной нам 10-ичной системе это были бы цифры от 0 до 9.

Представление числа a в форме с плавающей запятой или нормализованное представление означает его запись в виде

$$a = sign\ a \cdot r^p M = sign\ a \cdot r^p \cdot \left(\frac{b_1}{r} + \frac{b_2}{r^2} + ...\right)_r$$

где $sign\ a$ - знак числа a, p - порядок числа (целое); M мантисса числа a, причем $1/r < 1$, т.е. первая r—ичная цифра в записи мантиссы b_1 не равна нулю.

В современных ЭВМ в качестве основания системы счисления выбирается двойка- $r = 2$.

$$a = sign\ a \cdot 2^p \left(\frac{b_1}{2} + \frac{b_2}{2^2} + ... + \frac{b_t}{2^t} + \frac{b_{t+1}}{2^{t+1}} + ...\right)_2 \tag{6}$$

В двоичной системе цифры b могут принимать значения 0 и 1, при чем при представлении мантиссы в нормализованном виде $b_1 = 1$.

Число, представленное в виде (6), является бесконечной дробью. В t-разрядной ЭВМ хранятся разряды мантиссы только до t-ого порядка. Возникает вопрос, что делать с оставшимися разрядами. Мы рассмотрим простейший вариант усечения, отбрасывания всех порядков выше t. Тогда представление числа a в ЭВМ приводит нас к новому числу $\tilde{a}$. Точность представления числа a с помощью округленного числа a характеризуется относительной погрешностью округления

$$\delta_a = \frac{|a - \tilde{a}|}{|a|} - \quad \text{абсолютная относительная погрешность}$$

Оценим величину этой погрешности

$$\delta_a = \frac{|a-\tilde{a}|}{|a|} = \frac{2^p\left(\frac{b_{t+1}}{2^{t+1}} + \frac{b_{t+2}}{2^{t+2}} + ...\right)}{2^p\left(\frac{b_1}{2} + ...\right)} \leq \frac{\frac{1}{2^{t+1}}\left(1 + 1/2 + 1/4 + ...\right)}{1/2} = \frac{\frac{1}{2^{t+1}} \cdot 2}{1/2} = 2 \cdot 2^{-t}$$

Более точный способ округления дает для погрешности единичного округления вдвое меньшую оценку через машинный эпсилон

$$\frac{|a-\tilde{a}|}{|a|} = 2^{-t} = \varepsilon_m$$

Мы можем представления числа $\tilde{a}$ как результат действия некоторого оператора fl (fl. - floating point)

$$\tilde{a} = fl(a) = a(1+\varepsilon_a)$$

где $|\varepsilon_a| = \left|\frac{a-\tilde{a}}{a}\right| \leq \varepsilon_m$

Пример. Рассмотрим задачу о нахождении произведения n сомножителей

$$y_n = \prod_{i=1}^{n} z_i$$

Сформулируем алгоритм вычисления $\overline{A}$ следующим образом

$$\overline{A} = \begin{cases} y_k = z_k \cdot y_{k-1}, & k = 1, 2, ...n \\ y_0 = 1 \end{cases}$$

В результате реализации соответствующих вычислений y_{k-1} из-за погрешностей пришло к нам в виде $\tilde{y}_{k-1}$, тогда

$$\tilde{z_k} = fl(z_k \cdot \tilde{y}_{k-1}) = z_k \cdot \tilde{y}_{k-1}(1+\varepsilon_k) = \tilde{z}_k\tilde{y}_{k-1}$$

Тем самым мы получаем алгоритм $\tilde{A}$

$$\tilde{A} = \begin{cases} \tilde{y}_k = \tilde{z}_k \cdot \tilde{y}_{k-1}, & k = 1, 2, ...n \\ \tilde{y}_0 = \tilde{1} \end{cases}$$

Мы видим, что алгоритм $\tilde{A}$ — это есть алгоритм $\overline{A}$ на других возмущенных данных

$$\tilde{A} \to \overline{A}(\tilde{x})$$

2. Лекция 2. Интерполяция и приближение функций

Предположим, у нас есть некоторое множество функций F, и мы рассматриваем некоторую функцию $f(x) \in F$. Задача о нахождении значения функции в конкретной точке является достаточно трудоемкой. Тогда во множестве функции F выделим множество $G \subset F$ более "простых"функций $g(x)$, которые позволяют нам приближенно вычислить значение функции $f(x)$ в интересующих нас точках

$$g(x) \approx f(x)$$

Рассмотрим два варианта приближения функций:

1) Аппроксимационный(минимизационный) вариант. Рассматриваем, метрику порожденную соответствующей нормой. Тогда функция $g(x)$ ищется как

$$\min_G \|f(x) - g(x)\|$$

2) Интерполяционный вариант. Выбираем некоторое множество значений аргумента $\{x_k\}$ и ищем такую функцию $g(x)$, что

$$g(x_k) = f(x_k)$$

Если значения искомой функции $f(x)$ совпадает в некоторых точках с приближающей функцией $g(x)$, то мы вправе надеяться что и в других точках значения функций будут близки.

2.1. Постановка задачи интерполяции. Линейная интерполяция.

Сетка ω_n совокупность точек $\{x_i\}_{i=\overline{0,n}}$ заданных в области определения некоторой функции.

Замкнутая сетка $\overline{\omega_n}$ - граничные точки отрезка входят как начальные и конечные точки интерполяционной сетки.

Невырожденная сетка - ни один из узлов сетки не совпадает.

Пусть на отрезке $[a, b]$ задана замкнутая невырожденная сетка $\overline{\omega_n}$

$$a = x_0 \leq x_1 \leq x_2 \leq ... \leq x_n = b$$

и в ее узлах заданны значения функции

$$f(x_i) = f_i = y_i \quad i = \overline{0, n}$$

Обозначим кратко данную нам входную информацию как

$$\{f(x_i), x_i\}_{i=\overline{0,n}} \tag{7}$$

Наша задача построить интерполяционную функцию интерполянту $g(x)$, которая совпадает со значениями интерполируемой функции $f(x)$ в узлах сетки $\overline{w_n}$

$$g(x_i) = f(x_i); \quad i = \overline{0, n} \tag{8}$$

Мы ограничимся случаем линейной интерполяции, когда функция $g(x)$ представима в виде линейной комбинации

$$g(x) = \sum_{k=0}^{n} c_k \Phi_k(x)$$

Φ_k - известные, относительно легко вычислимые в каждой точке x линейно независимые функции; c_k - неизвестные коэффициенты. Причем индекс суммирования n в точности равен индексу n характеризующего входную информацию.

Из условия интерполяции (8) запишем условия на коэффициенты c_k:

$$\begin{cases} c_0\Phi_0(x_0) + c_1\Phi_1(x_0) + ... + c_n\Phi_n(x_0) = f(x_0) \equiv y_0 \\ c_0\Phi_0(x_1) + c_1\Phi_1(x_1) + ... + c_n\Phi_n(x_1) = f(x_1) \equiv y_1 \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ c_0\Phi_0(x_n) + c_1\Phi_1(x_n) + ... + c_n\Phi_n(x_n) = f(x_n) \equiv y_n \end{cases}$$

Мы пришли к СЛАУ относительно коэффициентов c_k. Разрешимость этой системы связана с характеристиками ее детерминанта.

Предположим, что система функций $\{\Phi_k(x)\}$ такова, что при любом невырожденном выборе узлов сетки $\overline{\omega_n}$ отличен от нуля определитель

$$\triangle(\Phi_0, \Phi_1, ..., \Phi_n) = \begin{vmatrix} \Phi_0(x_0) & \Phi_1(x_0) & \dots & \Phi_n(x_0) \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_0(x_n) & \Phi_1(n) & \dots & \Phi_n(x_n) \end{vmatrix} \neq 0$$

тогда такая система интерполяционных функций называется *чебышевской системой* интерполяционных функций.

Теорема 1. Для разрешимости задачи линейной интерполяции необходимо и достаточно чтобы система функций $\{\Phi_k(x)\}$ образовывала на $[a; b]$ чебышевскую систему интерполяционных функций.

Мы ограничимся рассмотрением только условиями интерполяции вида (7), разумеется, такие условия могут налагаться не только на саму функцию, но и на ее производные и т.д.

В качестве систем линейно независимых интерполяционных функций $\{\Phi_k(x)\}$ чаще всего выбирают степенные или полиномиальные функции $\Phi_k(x) = x^k$; тригонометрические функции $\Phi_k(x) = \begin{Bmatrix} cos(kx) \\ sin(kx) \end{Bmatrix}$ показательные функции $\Phi_k(x) = e^{kx}$ и другие. Мы ограничимся рассмотрением случая полиномиальной интерполяции.

2.2. Полиномиальная интерполяция

Пусть входной информацией является множество точек $y_k; x_k$. Искомая интерполяционная функция представляет собой многочлен $p_n(x)$

$$g(x) = \sum_{k=0}^{n} C_x x^k = C_0 + C_1 x + ... + C_n x^n$$

Условия интерполяции (8) приводят нас к системе линейных алгебраических уравнений да коэффициентов $\{c_k\}$

$$\begin{cases} c_0 + c_1 x_0 + c_2 x_0^2 + ... + c_n x_0^n = f_0 \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ c_0 + c_1 x_n + c_2 x_n^2 + ... + c_n x_n^n = f_n \end{cases}$$

Определитель этой системы (определитель Вандермонда) отличен от нуля на произвольной невырожденной сетке $\overline{\omega_n}$. Действительно, если раскрывать определитель по элементам последней строки, мы получим многочлен относительно x_n. С другой стороны, можно воспользоваться тем, что если в определителе есть две совпадающие строки он равен нулю. Тогда мы можем легко угадать корни нашего выражения, рассматривая определитель как функцию аргумента x_n. Например, $x_0 = x_n$, тогда первая и последняя строчка совпадают, значит определитель равен нулю, и это корень нашего уравнения. Так же поступаем и с x_1, и так далее до x_{n-1}. Далее рассматриваем определитель как функцию x_{n-1} и проводим те же действия, пока не появится последний сомножитель.

$$W = \begin{vmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{vmatrix} =$$

$$= (x_n - x_0)(x_n - x_1)\dots(x_n - x_{n-1})(x_{n-1} - x_0)(x_{n-1} - x_1)\dots(x_{n-1} - x_{n-2})\dots(x_1 - x_0) \neq 0$$

Так как сетка невырождена, ни один из сомножителей не совпадает и определитель не равен нулю, значит это система чебышевских функций. Тем самым задача однозначно разрешима.

Теорема 2. Интерполяционный многочлен $P_n(x)$ существует и единственен для любой невырожденной сетки.

2.3. Интерполяционный полином Лагранжа

При построении интерполяционного полинома $P_n(x)$ мы взяли в качестве системы функций $\Phi_k(x) = x^k$. С определенной точки зрения более удобной является система полиномов степени n, называемая базисом Лагранжа $\{l_k^{(n)}(x)\}$ и определенная из соображений: каждый $l_k(x)$ - полином n-ой степени, равный нулю во всех узлах сетки $\overline{\omega_n}$ кроме k-го, где он равен 1:

$$l_k^{(n)}(x_i) = \begin{cases} 1, & i = k \\ 0, & i \neq k \end{cases} \qquad k = \overline{0, n}$$

Нетрудно построить эти полиномы. Действительно, зная что из всех $n+1$ узлов полином не обращается в ноль только в k-ом (рис.2.2), мы можем указать все n корней многочлена.

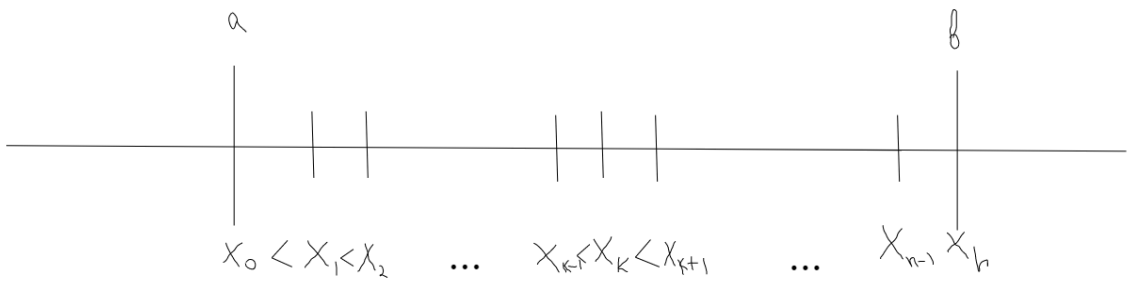

Рис. 2.1 – Разбиение отрезка [a,b]

$$l_k(x) = \frac{(x-x_0)(x-x_1)\dots(x-x_{k-1})(x-x_{k+1})\dots(x-x_n)}{(x_k-x_0)(x_k-x_1)\dots(x_k-x_{k-1})(x_k-x_{k+1})\dots(x_k-x_n)} = \prod_{\substack{i=0 \\ i\neq k}}^{n} \frac{x-x_i}{x_k-x_i}$$

Преобразуем базис $\{l_k(x)\}$. Введем в рассмотрение полином $(n+1)$-ой степени

$$\omega(x) = (x-x_0)(x-x_1)\dots(x-x_n) \tag{9}$$

Найдем производную $\omega(x)$

$$\omega'(x) = (x-x_1)\dots(x-x_n) + (x-x_0)(x-x_2)\dots(x-x_n) + \dots +$$

$$+(x-x_0)\dots(x-x_{k-1})(x-x_{k+1})\dots(x-x_n) + \dots + (x-x_0)\dots(x-x_{n-1})$$

Вычислим значения производной в точке x_k, тогда все слагаемые содержащие $(x-x_k)$ обратятся в ноль

$$\omega'(x) = (x_k-x_0)\dots(x_k-x_{k-1})(x_k-x_{k+1})\dots(x_k-x_n)$$

Тогда $l_k(x)$ преобразуется к виду

$$l_k^{(n)}(x) = \frac{\omega(x)}{(x-x_k)\omega'(x_k)}$$

и есть базис полиномов Лагранжа.

Построив эти базисные функции, мы видим, что $f_k l_k^n(x)$ во всех узлах сетки равно нулю, кроме k-го узла в котором равно f_k, поэтому

$$L_n(x) = \sum_{k=0}^{n} f_k l_k^{(n)}(x) = \sum_{k=0}^{n} f_k \frac{\omega(x)}{(x-x_k)\omega'(x_k)} \tag{10}$$

Полином $L_n(x)$ называют интерполяционным полиномом Лагранжа.

Заметим, что $L_n(x)$ – это то же самое что $P_n(x)$, а также $g(x)$, поскольку интерполянта единственна.

2.4. Интерполяционный полином Ньютона

Удобным представлением интерполяционного полинома для практических вычислений (особенно ручных) является запись интерполяционного полинома в виде интерполяционного полинома Ньютона.

Для этого введем в рассмотрение так называемые разделенные разности сеточной функции $\{f_i\}$. Определим их рекуррентно.

Для построения разделенной разности нулевого порядка нужен один узел:

$$f(x_i) = f(x)\Big|_{x_i}$$

Как мы видим, для разности нулевого порядка никакого нового обозначения не возникло

Разделенные разности первого порядка, построенные на двух узлах x_i, x_k определяются следующим образом

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

Разделенная разность второго порядка на узлах x_i, x_k, x_j определяется как первая разделенная разность от предыдущих разделенных разностей

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

Для разделенной разности порядка n нужно $n+1$ узлов. Построим по индукции разность k-го порядка

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

С помощью метода математической индукции доказывается, что разделенные разности можно представить следующим образом

$$f(x_0, x_1, \ldots, x_k) = \prod_{j=0}^{k} \frac{f(x_j)}{\prod_{\substack{l=0 \\ l \neq j}}^{k} (x_j - x_l)}$$

Отсюда следует вывод, что разделенные разности не зависят от способа упорядочения узлов, а зависят только от их количества.

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

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

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

разности	0-го	1-го	2-го	$\cdots$	$(n-1)$-го	n-го
x_0	$f(x_0)$			$\cdots$		
		$f(x_0,x_1)$		$\cdots$		
x_1	$f(x_1)$		$f(x_0,x_1,x_2)$	$\cdots$		
		$f(x_1,x_2)$		$\cdots$		
x_2	$f(x_2)$		$f(x_1,x_2,x_3)$	$\cdots$		
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$		
				$\cdots$	$f(x_0,\ldots,x_{n-1})$	
				$\cdots$		$f(x_0,\ldots,x_n)$
				$\cdots$	$f(x_1,\ldots,x_n)$	
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$		
x_{n-1}	$f(x_{n-1})$		$f(x_{n-2},x_{n-1},x_n)$	$\cdots$		
		$f(x_{n-1},x_n)$		$\cdots$		
x_n	$f(x_n)$			$\cdots$		
кол-во	$n+1$	n	$n-1$	$\cdots$	2	1

Рис. 2.2 – таблица значений функции

Далее мы будем рассматривать разделенные разности, составленные только по соседним узлам. Исходная таблица значений функции $\{f(x_i)\}$ позволяет построить следующие разделенные разности типа рис. 2.2.

Отметим особенности этих разделенных разностей. Если табличная функция сама по себе полином n-ой степени, т. е. $f(x) = p_n(x) \equiv p(x)$, то ее первая разделенная разность

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

есть полином $n-1$-ой степени, поскольку в числителе дроби стоит полином степени n, равный нулю при $x = x_0$, т. е. делящийся нацело на $(x - x_0)$.

Вторая разделенная разность

$$p(x, x_0, x_1) = \frac{p(x, x_0) - p(x_0, x_1)}{x - x_1}$$

является полиномом $n-2$-ой степени, поскольку в числителе полином $n-1$-ой степени, равный нулю в точке $x = x_1$.

Разделенная разность n-го порядка

$$p(x, x_0, \ldots, x_{n-1}) = const$$

является полиномом 0-го порядка, т.е константой.

Разделенная разность $n+1$-го порядка

$$p(x_0, x_1, \ldots, x_n) = \frac{p(x, x_0, \ldots, x_{n-1}) - p(x_0, \ldots, x_n)}{x - x_n} \equiv 0$$

тождественно равна нулю, поскольку в числителе стоят разделенные разности n-го равные одной и той же константе.

Вернемся к разности n-го порядка

$$p(x, x_0, \ldots, x_{n-1}) = p(x_0, x_1, \ldots, x_n) + \underbrace{(x - x_n)p(x, x_0, \ldots, x_n)}_{\equiv 0}$$

Осуществляя обратную подстановку (рекуррентно) найдем

$$p(x) = p(x_0)+(x-x_0)[p(x_0,x_1)+(x-x_1)[p(x_0,x_1,x_2)+(x-x_2)[\ldots(x-x_{n-1})p(x_0,\ldots,x_n)] =$$

$$p(x_0) + (x - x_0)p(x_0, x_1) + (x - x_0)(x - x_1)p(x_0, x_1, x_2) + \ldots$$
$$\ldots + (x - x_0)(x - x_1)\ldots(x - x_{n-1})p(x_0, \ldots, x_n)$$

Осталось найти представление не произвольного многочлена $p_n(x)$, а интерполяционного многочлена $P_n(x)$ для функции $f(x)$, тогда $p(x_i) = f(x_i)$, и мы получим явную форму записи интерполяционного многочлена

$$P_n(x) \equiv f(x_0) + (x - x_0)f(x_0, x_1) + \ldots + (x - x_0)\ldots(x - x_{n-1})f(x_0, \ldots, x_n) \quad (11)$$

в виде интерполяционного многочлена Ньютона.

2.5. Погрешность полиномиальной интерполяции

Итак, мы нашли основные представления интерполяционной функции $g(x)$. Напомним что наша задача состоит в том, чтобы построить функцию $g(x)$, с помощью которой можно приблизить значения функции $f(x)$ в точках, не принадлежащих $\overline{\omega_n}$ (ведь в точках $x \in \overline{\omega_n}$ задача интерполяции в точности равна $f(x) = g(x)$).

Остановимся теперь на вопросе о погрешности полиномиальной интерполяции. Запишем вспомогательную функцию

$$\varphi(z) = f(z) - P_n(z) - A\omega(z),$$

где $A = const$, $\omega(z)$ - введенный ранее полином (9) $n+1$-ой степени.

Определим $const$ из того условия, чтобы в произвольной фиксированной точке $x \notin \overline{\omega_n}$, выполнялось равенство $\varphi(x) = 0$. Тогда

$$A = \frac{f(x) - P_n(x)}{\omega(x)} \quad \text{при } x \notin \overline{\omega_n}$$

Итак, мы добились того, что в конкретной точке x функция $\varphi(x)$ равна нулю. Предположим, что наша функция $f(x)$ имеет непрерывные до $(n+1)$ порядка включительно производные на $[a; b]$, то есть $f(x) \in C^{(n+1)}[a, b]$. Тогда функция $\varphi(z)$ - непрерывна на $[\tilde{a}; \tilde{b}]$, где $\tilde{a} = \min(x, a)$, $\tilde{b} = \max(x, b)$. Мы добавили к узлам нашей сетки $x_0, x_1, \ldots, x_n$ еще один узел x, тем самым получили $n + 2$ узла. На интервале $[\tilde{a}; \tilde{b}]$ функция $\varphi(z)$ обращается в ноль в $n + 2$ точках, и по теореме Ролля можно утверждать, что существует $n + 1$ внутренняя точка на $[\tilde{a}; \tilde{b}]$ где $\varphi'(z) = 0$. Аналогично существует n точек, где $\varphi''(z) = 0$, и т.д.

Следовательно, существует одна точка $\xi \in (\tilde{a}; \tilde{b})$, в которой

$$\varphi^{(n+1)}(z)\Big|_{\xi} = 0$$

С другой стороны

$$\varphi^{(n+1)}(\xi) = f^{(n+1)}(\xi) - A(n+1)! = 0$$

($(n+1)$-ая производная от полинома n-ой степени $P_n(z)$ тождественно равна 0).

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

Полученные соотношения позволили нам установить формулу

$$f(x) - P_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega(x) \quad \xi \in (\tilde{a};\tilde{b})$$

Это и есть исходное выражение, которое позволяет получить оценку погрешности интерполяции. Имеем

$$|f(x) - P_n(x)| \le \frac{M_{n+1}}{(n+1)!}|\omega(x)| \tag{12}$$

где $M_{n+1} = \max_{[\tilde{a};\tilde{b}]} |f^{(n+1)}(x)|$

Дальнейшее использование оценки (12) связано с изучением характера поведения $|\omega(x)|$ при произвольном расположении узлов интерполяции, что достаточно сложно и громоздко. Ограничимся наиболее часто рассматриваемым на практике случаем:

1) Равномерной сетки $\overline{\omega_n}$ с постоянным шагом $h = \frac{b-a}{n}$, $x_k = x_0 + kh$

2) Узлы интерполяции на этой сетке выбраны подряд.

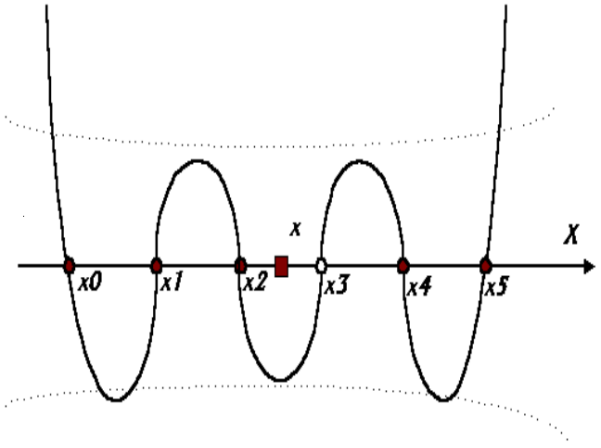

Рис. 2.3 – примерный вид многочлена при n=5

Для наглядности выберем $n = 5$. Тогда полином $\omega(x)$ имеет следующий вид

$$\omega(x) = (x - x_0)(x - x_1)\dots(x - x_5)$$

Эскизно наблюдая за поведением функции, можно сделать вывод что, вне отрезка $[a, b]$ выбирать точку x крайне невыгодно, потому что оценка погрешности будет явно завышена из-за быстрого роста функции $|\omega(x)|$. Говорят, что *экстраполяция ненадежна* при $x \in [a, b]$. Если выбрать точку x в центре интерполяционных узлов, то значение $|\omega(x)|$ близко к экстремуму. Сравнительно просто оценка погрешности интерполяции проводится в случае нечетного n. Пусть при этом рассматриваемое x

находится в центральном интервале. На этом интервале экстремум $\omega(x)$ достигается точно в середине центрального интервала сетки $\overline{\omega_n}$, и именно это значение входит в максимум $|\omega(x)|$:

$$|f(x) - P_n(x)| \leq \frac{M_{n+1}}{(n+1)!}|\omega(x)| \leq C(n)\frac{M_{n+1}}{(n+1)!}h^{n+1} \tag{13}$$

Замечание. Неравенство (13) существует не всегда. Оно справедливо при рассмотрении равномерной сетки и центрирования сетки нужным образом.

Если мы рассматриваем оценку (13) как функцию h (фиксируем степень многочлена и рассматриваем сетки, характеризующиеся различными интервалами h), то погрешность интерполяции будет убывать как величина порядка $O(h^{n+1})$. Поэтому говорят, что интерполяционный многочлен $P_n(x)$ обеспечивает $(n+1)$-ый порядок точности интерполяции и интерполяция имеет погрешность $O(h^{n+1})$

2.6. Сходимость интерполяционного многочлена

Рассмотрим на отрезке $[a, b]$ последовательность сеток (таких последовательностей бесконечно много)

$$\{\overline{\omega_n}\} = \{\overline{\omega_0}, \overline{\omega_1}, \ldots, \overline{\omega_n}\}$$

Этой последовательности сеток соответствует последовательность интерполяционных многочленов $\{P_n(x)\}$. Теперь нам нужно рассмотреть вопрос о сходимости $\{P_n(x)\}$ при стремлении числа узлов n к бесконечности.

1) Многочлен $P_n(x)$ сходится поточечно к функции $f(x)$ на $[a, b]$

$$P_n(x) \to f(x) \quad \text{для } \forall x \in [a, b]$$

2) Многочлен $P_n(x)$ сходится равномерно к функции $f(x)$ на $[a, b]$, если

$$\lim_{n\to\infty} \sup_{[a,b]} |f(x) - P_n(x)| = 0$$

обозначается как $P_n(x) \rightrightarrows f(x)$

3) Многочлен $P_n(x)$ сходится в среднем к функции $f(x)$ на $[a, b]$, если

$$\lim_{n\to\infty} \int_a^b (f(x) - P_n(x))^2 \, dx = 0$$

Подчеркнем еще раз, что последовательность сеток $\overline{\omega_n}$ фиксирована при рассмотрении соответствующих пределов.

С практической точки зрения, сходимость интерполяции можно изучать следующим образом:

1) Сохраняя степень интерполяционного полинома n, уменьшая шаг сетки. ($h \to 0, n = const$)

2) Сохранять шаг сетки, увеличивать число используемых узлов интерполяции на $[a, b]$, то есть увеличивать степень интерполяционного многочлена: $n \to \infty, h = const$.

1) Уменьшение шага сетки $\overline{\omega_n}$. Как мы уже отмечали, погрешность метода при интерполяции многочленом $P_n(x)$ есть, согласно (6), величина порядка $O(n+1)$, т.е. $f(x) - P_n(x)$ неограниченно убывает при $h \to 0$, при этом интерполяционный многочлен сходится к $f(x)$ в некотором смысле "равномерно".

$$|\omega(x)| = \left| \prod_{i=0}^{n} (x - x_i) \right| \leq (nh)^{n+1}$$

Тогда, равномерно по x, можно в формуле (13) провести оценку

$$|f(x) - P_n(x)| \leq \frac{M_{n+1}}{(n+1)!}(nh)^{n+1} = O(h^{n+1})$$

Увеличения числа узлов. Нужно сразу же оговориться, что увеличение числа узлов, то есть степени интерполяционного многочлена, не всегда целесообразно, так как:

а) неизвестно как быстро растет оценка максимума модуля производной M_{n+1} с ростом ее порядка;

б) у функции $f(x)$ может быть лишь ограниченное число непрерывных производных.

В рассмотрении этого случая ограничимся рассмотрением лишь отдельных примеров.

Пример 1.Пример Берштейна

Для функции $y = |x|$, интерполяция на равномерной на отрезке $[-1, 1]$ сетке $h = const$ не дает поточечной сходимости ни в одной точке , кроме $x \in \{-1, 0, 1\}$

$$P_n(x) \not\to f(x) \quad \forall x \in \{[-1, 1] \backslash \{-1, 0, 1\}\}$$

Тем самым интерполяционный процесс расходится.

Пример 2

Пусть функция $f(x)$ является бесконечно дифференцируемой на $[a, b]$ и для нее существует оценка $M_n = m^n$, тогда можно получить равномерную сходимость

$$\max_{[a,b]} |f(x) - P_n(x)| \leq \frac{(m(b-a))^{n+1}}{(n+1)!} < \varepsilon$$

Дробь стремится к нулю, так как факториал возрастает быстрее, чем степенная функция.

Такие целые функции являются для нас удобными, но их запас не столь "велик"для практических целей.

Для интерполяции непрерывных функций приведем формулировки двух теорем.

Теорема 3 (Фабера)

Для любой последовательности сеток найдется непрерывная на $[a, b]$ функция, для которой нет равномерной сходимости.

$$\forall\{\overline{\omega_n}\} \ \exists f(x) \in C[a, b] \quad \text{что} \quad P_n(x) \not\Rightarrow f(x)$$

Теорема 4 (Марцинкевича)

Для любой непрерывной на отрезке $[a, b]$ функции найдется такая последовательность сеток, что интерполяционный многочлен сходится к этой функции.

$$\forall f(x) \in C[a,b] \ \exists\{\overline{\omega_n}\} \quad \text{что} \quad P_n(x) \Rightarrow f(x)$$

Эти теоремы еще раз подчеркивают, что сходимость интерполяционного многочлена существенно связана с расположением узлов сетки.

Общий вывод: В практике вычислений избегают использования метода увеличения степени интерполяционных полиномов. Вместо этого для интерполяции $f(x)$ на большом отрезке стараться разбить его на частичные отрезки и на каждом частичном отрезке использовать интерполяционные свойства для многочленов не слишком большого порядка. То есть использовать кусочно-полиномиальную интерполяцию (сплайн-интерполяция).

3. Лекция 3. Сплайн-интерполяция

Сплайном $S_n^\alpha(x)$ порядка n и дефектом α называется кусочно-полиномиальная функция непрерывная на отрезке $[a,b]$ вместе со своими производными до порядка $n-\alpha$ и обладающая свойством: $\forall x \in \triangle_x = [x_{k-1}, x_k]$. Соответствующий сплайн является многочленом степени n, $S_n^\alpha(x) = P_n$

Определим количество свободных параметров, характеризующих сплайн. На каждом отрезке это многочлен n-го порядка, который определяется своими коэффициентами, их количество равно $n+1$. Если число интервалов равно M, тогда всего нам требуется $M(n+1)$ параметр. Но также нам нужно записать непрерывность производных до $n-\alpha$-го порядка во внутренних узлах, таких $M-1$ штука и в самой функции. Тогда количество свободных параметров описывающих равно

$$M(n+1) - (M-1)(n-\alpha+1)$$

Мы ограничимся рассмотрением распространенного частного случая сплайна третьего порядка или кубического сплайна.

3.1. Кубический сплайн

Рассмотрим кубический сплайн с дефектом равным 1, s_3^1, количество характеризующих его параметров равно

$$4M - (M-1)3 = M+3$$

Наша цель - решить задачу интерполяции для функции $f(x)$. Построим интерполянту $g(x)$, в качестве которой мы хотим предложить сплайн. Пусть задана интерполяционная сетка

$$\overline{\omega_n} = \{x_0, x_1, \ldots, x_n\}$$

Подчеркнем, что интерполяционная сетка и сетка, определяющая сплайн, – в общем случае не одно и то же. Нам необходимо, чтобы сплайн удовлетворял $n+1$-му условию интерполяции

$$S(x_k) = f_k \quad \text{где} \quad k \in \overline{0,n}$$

Выделим условно в свободных параметрах сплайна две группы, в первой будут $m+1$ параметр, во второй оставшиеся 2 параметра. Тогда первая группа параметров будет удовлетворять $n+1$-му условию интерполяции, а вторая описывать поведение сплайна на границе $[a,b]$ (граничные условия). Теперь мы можем согласовать две сетки. В качестве узлов, определяющих сплайн, нужно использовать узлы интерполяционной сетки, положив n равным m.

Задача. В качестве самостоятельного упражнения по аналогии придумайте, как нужно согласовать сетку и расположить узлы для сплайна $S_3^2(x)$.

Теорема 5

Для любой дважды непрерывно дифференцируемой на $[a,b]$ функции $f(x) \in C^2[a,b]$, удовлетворяющей интерполяционной таблице $f(x_i) = y_i$ с однородными

краевыми условиями ($f''(a) = f''(b) = 0$), где $x_i,\ i = \overline{0,n}$ интерполяционный набор точек, интерполяционный сплайн $S_3^1(x) \equiv S_x$ доставляет минимум следующему функционалу

$$\Phi[f] = \int_a^b (f''(x))^2\,dx$$

Или, другими словами, нужно показать, что для любой функции $f(x) \in C^2[a,b]$

$$\Phi[f] \geq \Phi[S]$$

Доказательство.
Рассмотрим следующую разность

$$\Phi[f] - \Phi(s) = \int_a^b (f'')^2 - (s'')^2\,dx = \underbrace{\int_a^b (f'' - s'')^2\,dx}_{\geq 0} + \underbrace{\int_a^b (2f''s'' - 2s'')}_{I_2}$$

Рассмотрим отдельно второй интеграл I_2, представив его в виде суммы отдельных интервалов и интегрируя по частям

$$2\int_a^b (f'' - s'')s''\,dx = 2\sum_{k=1}^n \int_{x_{k-1}}^{x_k} (f'' - s'')s''\,dx =$$

$$= 2\left[\sum_{k=1}^n \underbrace{(f' - s')s''}_{\varphi(x)}\Big|_{k-1}^{k} - \sum_{k=1}^n \int_{x_{k-1}}^{x_k} (f' - s')s'''\,dx\right]$$

Рассмотрим подробнее, что собой представляется функция, которую мы обозначили $\varphi(x)$

$$\sum_{k=1}^n \varphi(x) = \varphi_1 - \varphi_0 + \varphi_2 - \varphi_1 + \ldots + \varphi_n - \varphi_{n-1} = \varphi_1 - \varphi_n$$

Наш сплайн, также как и функция $f(x)$, удовлетворяет однородным краевым условиям, поэтому

$$\sum_{k=1}^n \varphi(x) = 0$$

Теперь рассмотрим отдельно выражение

$$\sum_{k=1}^n \int_{x_{k-1}}^{x_k} (f' - s')s'''\,dx = \sum_{k=1}^n s'''(f - s)\Big|_{x_{k-1}}^{x_k} = 0$$

Так сплайн на частичном отрезке является кубическим многочленом, то его производная третьей степени является константой, а значит мы можем вынести ее за знак интеграла. Так как сплайн интерполяционный, то в точках x_k он принимает значение y_i, также как и функция $f(x)$. Тем самым весь интеграл I_2 равен нулю, а значит

$$\Phi[f] - \Phi[S] \geq 0 \quad \forall f$$

что и требовалось доказать.

Вернемся к решению задачи интерполяции через построение нормального (с граничными условиями $s''(x_0) = s''(x_n) = 0$) кубического сплайна $S_3^1(x) \equiv s(x)$. На k-ом частичном отрезке $s(x)$ можно записать, используя формулу Тейлора, как

$$s(x)\big|_{\triangle_k} = a + bx + cx^2 + dx^3 = \tilde{a} + \tilde{b}(x - x_k) + \tilde{c}\frac{(x - x_k)^2}{2!} + \tilde{d}\frac{(x - x_k)^3}{3!}$$

Нетрудно показать, что функция s'' является линейной на частичном отрезке и непрерывной в силу свойств сплайна(рис.3.1).

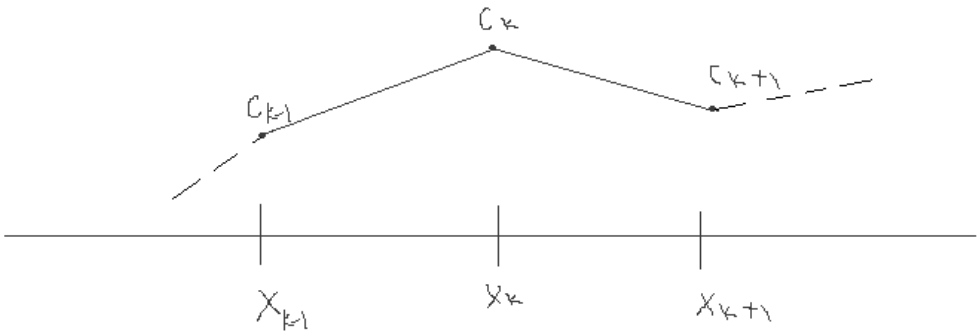

Рис. 3.1 – примерный вид функции s”

А на всем отрезке $[a, b]$ кусочно-линейной и кусочно-непрерывной. Коэффициенты $\tilde{C}_k$ являются значениями второй производной в узлах x_k. Пока коэффициенты сплайна для нас неизвестны, но мы покажем, что их можно вычислить их через значения коэффициентов второй производной сплайна. Проинтерполируем функцию $s''(x)$, считая известными коэффициенты C_k.

$$s''(x) = L_1(x) = \frac{C_{k-1}(x - x_k)}{x_{k-1} - x_k} + \frac{C_k(x - x_{k-1})}{x_k - x_{k-1}}, \quad x \in \triangle_k$$

Если ввести обозначения для шагов сетки h_k и частичных отрезков $\triangle_k$, то представление $s''(x)$ окажется следующим:

$$s''(x) = -\frac{C_{k-1}}{h_k}(x - x_k) + \frac{C_k}{h_k}(x - x_{k-1}) \tag{14}$$

Интегрируя выражение (14), найдем $s'(x)$:

$$s'(x) = -\frac{C_{k-1}}{h_k}\frac{(x - x_k)^2}{2} + \frac{C_k}{h_k}\frac{(x - x_{k-1})^2}{2} + B_k$$

И, интегрируя еще раз, найдем $s(x)$:

$$s(x) = -\frac{C_{k-1}}{6h_k}(x - x_k)^3 + \frac{C_k}{6h_k}(x - x_{k-1})^3 + B_k x + A_k$$

Для того, чтобы удовлетворить условиям интерполяции, нам необходимо, чтобы значения $s(x)$ в точках x_k совпадали с y_k. Проведем соответствующие подстановки

$$\begin{cases} s(x_{k-1}) = c_{k-1}\frac{h_k^2}{6} + B_k x_{k-1} + A_k = y_{k-1} \\ s(x_k) = c_k\frac{h_k^2}{6} + B_k x_k + A_k = y_k \end{cases} \tag{15}$$

Теперь запишем условие непрерывности s' слева и справа от точки x_k на интервале $\triangle_k$ и $\triangle_{k+1}$

$$s'(x_k) = C_k\frac{h_k}{2} + B_k = -\frac{C_k}{h_{k+1}}\frac{h_{k+1}^2}{2} + B_{k+1} = s'(x_{k+1}) \tag{16}$$

Коэффициенты B_k и B_{k+1} найдем из системы (15), вычитая первое уравнение из второго

$$B_k = \frac{y_k - y_{k-1}}{h_k} - \left((C_k - C_{k-1})\frac{h_k}{6}\right)$$

Такая же формула будет для B_{k+1}, но со сдвинутыми индексами. Вернемся к выражению (16) и нахождению условий непрерывности

$$\begin{cases} C_{k-1}h_k + 2(h_k + h_{k+1})C_k + C_{k+1}h_{k+1} = 6\left(\frac{y_{k+1}-y_k}{h_{k+1}} - \frac{y_k - y_{k-1}}{h_k}\right), \quad k = 1, \dots, n-1 \\ C_0 = 0 \\ C_n = 0 \end{cases}$$

Эти уравнения дают нам СЛАУ с трехдиагональной матрицей. Запишем ее в следующем виде

$$A\begin{pmatrix} C_0 \\ \vdots \\ C_n \end{pmatrix} = (f)$$

Это матрица с диагональным преобладанием $|a_{ii}| > \sum_{j\neq i}|a_{ij}|$. По теореме Гершгорина собственные значения такой матрицы $\lambda(A)$ принадлежат объединению так называемых кругов Гершгорина $\bigcup_i |z - a_{ii}| < \sum_{j\neq i}|a_{ij}|$. Из этой теоремы следует $\lambda = 0$ не является собственным значением матрицы с диагональным преобладанием, то есть матрица не вырождена, а значит решение существует и единственно. Тем самым конструктивно показано, что нормальный кубический интерполяционный сплайн существует и единственен.

В чем состоит удобство и неудобство решения задач интерполяции с помощью кусочно-полиномиальных функций с точки зрения аппроксимации? Сформулируем теорему, устанавливающую характер сходимости интерполяционного сплайна.

Теорема 6

Пусть $f(x) \in C^4[a,b]$, и $\overline{\omega_n}$ – равномерная сетка $h = const$, тогда справедлива оценка

$$\|f(x) - s_3'(x)\|_{c[a,b]} = \max_{[a,b]}|f(x) - s_3'(x)| \leq C_0 M_4 h^4$$

$$||f'(x) - s'(x)||_{c[a,b]} \leq C_1 M_4 h^3$$

$$||f''(x) - s''(x)||_{c[a,b]} \leq C_2 M_4 h^2$$

где $M_4 = \sup_{[a,b]} |f^{(4)}(x)|$.

Тем самым сплайн-интерполяция выгодно отличается от полиномиальной интерполяции сходимостью и устойчивостью вычисления интерполяционного сплайна $s(x)$.

4. Лекция 4. Аппроксимация функций

1) Если мы ставим задачу построения функции $F(x) \in \mathcal{F}$, которая будет приближать значения функции $f(x)$, если мы говорим о приближение с точностью ε:

$$||f(x) - F_\varepsilon(x)|| < \varepsilon$$

2) Если мы ставим задачу о наилучшем приближении или приближению с заданной точностью:

$$||f(x) - \overline{F}(x)|| = \inf_{\mathcal{F}} ||f(x) - F(x)||$$

Мы ограничимся рассмотрением случая построения аппроксиманты для специального случая, когда соответствующая функции ищется в линейной оболочке удобных с этой точки зрения функций. Будем считать, что из $\mathcal{F}$ представляют собой функции интегрируемые с квадратом весом $\rho > 0$ на отрезке $[a, b]$

$$\int_b^a f(x)g(x)\rho\, dx = (f(x), g(x))$$

Это пространство $L_{2,p}[a, b]$ квадратичных суммируемых функций с весом ρ на соответствующем отрезке $[a, b]$

В курсе математического анализа мы рассматривали соответствующие пространства и задачи построения рядов Фурье по системе ортогональных функций. В таких пространствах существуют системы ортонормированных функций $\{\varphi_n(x)\}_{n=\overline{1,\infty}}$, таких что

$$(\varphi_k, \varphi_p) = \delta_{k,p}$$

. Будем ставить задачу аппроксимации, считая что функция $F(x)$, аппроксимирующая функцию $f(x)$, является элементом линейной оболочки порожденной функциями $\{\varphi_n(x)\}_{n=\overline{1,N}}$

$$F(x) \in Lin(\varphi_1, \ldots, \varphi_N) = \left\{ \varphi \in L_2[a, b]; \quad \varphi(x) = \sum_{k=1}^{N} C_k \varphi_k(x) \right\}$$

В таком случае можно говорить о задаче аппроксимации обобщенным многочленом по системе функций $\{\varphi_k(x)\}$.

$$F_N(x) = \sum_{k=1}^{N} C_k \varphi_k(x); \quad f_k = (f, \varphi_k)$$

4.1. Существование и единственность наилучшего среднеквадратичного приближения

Рассмотрим среднеквадратичное отклонение

$$\delta^2 = ||f - F||^2 = (f - F, f - F) = (f, f) - 2(f, F) + (F, F) = ||f||^2 + 2(f, F) + ||F||^2$$

Представление функции $F(x)$ записано по системе функций $\{\varphi_k(x)\}$, поэтому

$$\delta^2 = ||f||^2 - 2\left(f, \sum_k C_k\varphi_k\right) + \sum_{k,p} C_kC_p(\varphi_k, \varphi_p) = ||f||^2 - 2\sum_k C_kf_k + \sum_k C_k^2 =$$

$$= ||f||^2 + \sum_{k=1}^{N}(c_k - f_k)^2 - \sum_{k=1}^{N} f_k^2$$

Трансформировав таким образом величину среднеквадратичного отклонения и исходя из того, что коэффициенты Фурье функции f зафиксированы, так как посчитаны по системе функций $\{\varphi_k(x)\}$, и выражение $c_k - f_k$ больше нуля, мы можем сделать следующие выводы:

1) Решение задачи о нахождении минимального значения среднеквадратичного отклонения существует и единственно, ответ дается отрезком ряда Фурье по функциям $\{\varphi_k(x)\}$

2) При $c_k = f_k$, т.е на функции

$$\overline{F}_N(x) = \sum_{k=1}^{N} f_k \cdot \varphi_k(x) = \sum_{k=1}^{N}(f, \varphi_k)\varphi_k(x)$$

3) Минимальная величина среднеквадратичного отклонения равна

$$\delta_N^2 = ||f||^2 - \sum_{k=1}^{N} f_k^2$$

Тем самым мы показали, что если мы хотим решать задачи аппроксимации и добиваться наименьшего возможного значения среднеквадратичной погрешности в нашей аппроксимации, то мы должны пользоваться отрезком ряда Фурье по системе функций $\{\varphi_k(x)\}$.

Рассмотрим *равенство Парсеваля-Стеклова* являющееся, необходимым и достаточным условием замкнутости системы функций

$$\sum_{k=1}^{\infty} f_k^2 = ||f||^2$$

тогда

$$\delta_N^2 = \sum_{k+1}^{\infty} f_k^2 \to 0$$

Значит можно утверждать, что $\forall\varepsilon$ существует $N(\varepsilon)$, такое что для любого $N > N(\varepsilon)$

$$\delta_N^2 < \varepsilon^2$$

Тем самым можно решить задачу о среднеквадратичной аппроксимации с любой наперед заданной точностью.

Если система функций $\{\varphi_k(x)\}$ не ортогональна, но линейна независима, то выкладки отчасти усложняются

$$\delta^2 = ||f - F||^2_{L_{2,p}[a,b]} = \left(f - \sum_{k=1}^{N} C_k\varphi_k(x), f - \sum_{p=1}^{N} C_p\varphi_p(x)\right) = \Phi(C_1, \ldots, C_N)$$

Тем самым задача о построении наилучшего квадратичного приближения – это задача о минимизации функции $\Phi(C_1, \ldots, C_N)$.

$$\frac{\partial\Phi}{\partial C_m} = 2\left(-\varphi_m(x), f - \sum_{p=1}^{N} C_p\varphi_p(x)\right) = 0$$

Коэффициент 2 возникает из симметрии скалярного произведения. Мы получили СЛАУ для определения $\{c_k\}$.

$$\sum_{p=1}^{N} C_p(\varphi_m, \varphi_p) = (f, \varphi_m) \tag{17}$$

Ее определитель - это определитель Грама линейно независимой системы функций $\{\varphi_k(x)\}$. Он строго больше нуля

$$det||(\varphi_m, \varphi_p)|| = G(\varphi_1, \ldots, \varphi_N) > 0$$

Это означает, что существует и единственен набор коэффициентов $\overline{C_1}, \ldots, \overline{C_N}$, являющийся решением системы (17). С помощью этого набора мы строим наилучшее среднеквадратичное приближение

$$\overline{F}_N(x) = \sum_{k=1}^{N} \overline{C_k}\varphi_k(x)$$

В отличие от рассмотренного ранее случая, есть некоторые замечания

1) Изменение индекса N, то есть количества функций, участвующих в среднеквадратичной аппроксимации, приводит к изменению всей матрицы системы 17 и изменению коэффициентов C_k. В случае же, когда система функций $\{\varphi_k\}$ была ортонормирована или ортогональна, единожды вычисленный коэффициент $C_k = f_k$ не изменялся.

2) С ростом N определитель $G(\varphi_1, \ldots, \varphi_N)$ стремится к нулю, и система (17) становится плохо обусловленой, что приводит к дополнительным трудностям при решении СЛАУ.

Рассмотренная постановка о наилучшем среднеквадратичном приближение распространяется на случай, когда входная информация представлена не в виде функции $f(x)$, но и в том случае, когда мы имеем дело с дискретным набором ее значений. Но, в отличие от задач интерполяции, мы не требуем точного совпадения значений функций $F(x)$ и $f(x)$ на некотором наборе точек. В качестве интерпретации можно предложить экспериментальные данные, полученные с некоторой погрешностью.

4.2. Метод наименьших квадратов(МНК)

Пусть мы имеем следующую входную информацию $\{x_i; f(x_i) = f_i = y_i\}_{i=\overline{0,n}}$. Спроецируем систему функций $\{\varphi_k(x)\}_{k=\overline{1,N}}$ на дискретном множестве данных нам точек, получим множество векторов $\{\varphi_k(x_p)\}$, где $p = \overline{1,n}$, которые образуют это конечномерное $n+1$-мерное евклидово пространство столбцов $R^{n+1} = \mathcal{H}$. В данном пространстве скалярное произведение двух функций имеет вид

$$(\vec{x}, \vec{y})_{\mathcal{H}} = \sum_{p=0}^{n} x_p y_p \rho_p, \quad \vec{x}, \vec{y} \in \mathcal{H} \tag{18}$$

Если ρ тождественно равно единице, то выражение (18) равно просто сумме произведений одноименных координат. Рассмотрим задачу о линейной аппроксимации с помощью этих векторов входной информации.

$$\delta_N^2 = ||f - F_N||_{\mathcal{H}}^2 = \sum_{p=0}^{n} \left(f(x_p) - \sum_{k=1}^{N} C_k \varphi_k(x_p) \right)^2 \rho_p = \Phi(C_1, \ldots, C_N)$$

Задача построения наилучшего среднеквадратичного приближения снова сводится задаче минимизации функции $\Phi(C_1, \ldots, C_N)$. Сразу выпишем нормальную систему

$$\sum_{k=1}^{N} C_k(\varphi_k, \varphi_p) = (f, \varphi_p) \tag{19}$$

Определитель этой системы также является определителем Грама, но совокупности векторов. Распишем отдельно что представляет собой скалярное произведение выражения (19)

$$(\varphi_k, \varphi_p) = \sum_{i=0}^{n} \varphi_k(x_i) \varphi_p(x_i) \rho_i$$

Определитель Грама $G(\rho_0, \ldots, \rho_N) \neq 0$ для системы линейно-независимых сеточных функций, следовательно, решение задачи (19) существует и единственно. Также как и для решения задач аппроксимации удобнее использовать системы ортонормированных функций.

4.3. Обработка экспериментальных кривых методом НК

МНК широко используется в обработке экспериментальных кривых, т.е. таких кривых, точки которых измерены с известной погрешностью ε_i - $\{x_i; f(x_i) = f_i = y_i; \varepsilon_i\}$.

$$\delta_N^2 = \sum_{k=0}^{n} \rho_k [f(x_k) - F_N(x_k)]^2 \tag{20}$$

В таком случае обычно весу ρ_k придают смысл точности измерения отдельной точки, полагая

$$\rho_k = \frac{1}{\varepsilon_k^2}$$

Тогда аппроксимирующая кривая будет проходить "ближе"к точкам с большим весом (где выше точность), ибо каждое слагаемое в $\delta_N^2 = ||f - F_N||^2$ заведомо не превосходит ε^2 и в произведении

$$\rho_k(f(x_k) - F_N(x_k))^2$$

второй сомножитель должен быть существенно меньше для получения той же величины результата. Нарисуем график аппроксимации некоторой физической величины по исходным данным с относительно большими погрешностями (рис. 4.1).

Среднеквадратичное отклонение для интерполяционного многочлена на этой сетке в точности равно нулю, тем не менее с физической точки зрения информация о характере изменения функции передано неверно. При обработке экспериментальных данных и построения функции нужно контролировать параметры N, n и δ_N^2.

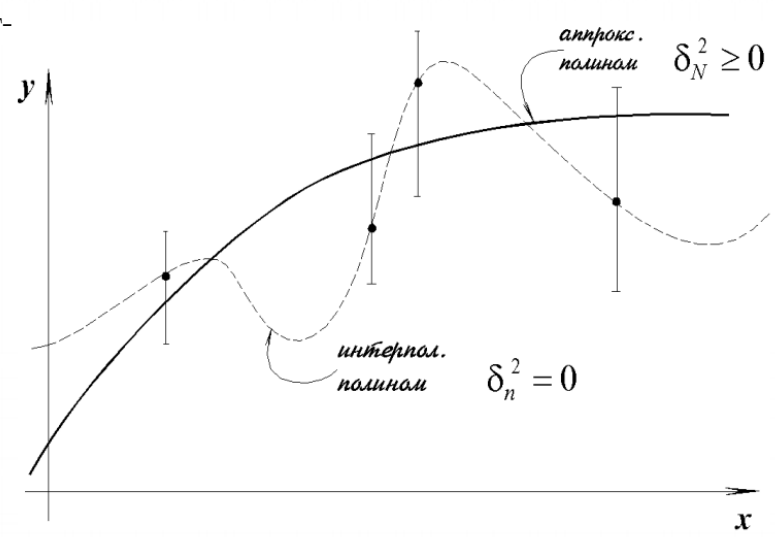

Рис. 4.1 – приближение функции разными способами

Если выбирать $N \geq n$, то в запас функций будет входить интерполяционный многочлен, и характер функциональной зависимости будет передан неправильно. Хорошее сглаживание эксперимента будет при $N \ll n$. При этом величина δ_N^2 должна быть порядка погрешности входных данных, $\delta_N \approx \varepsilon = \max_{\{i\}} \varepsilon_i$. Если нам удается выполнить эти условия, то можно считать, что среднеквадратичная аппроксимация по методу наименьших квадратов решена и функция $F_N(x)$, удовлетворяющая условию (20), является подходящим среднеквадратичным приближением для метода МК.

Эти рассуждения объясняют тактику подбора в построении функции $F_N(x)$. Мы начинаем с небольшого числа N – например, одна функция. Понятно, что с помощью одной функции вряд ли получится хорошо передать зашифрованную во входных данных функцию $f(x)$. Поэтому мы ожидаем, что погрешность δ_1^2 будет существенно больше, чем ε. Тогда мы будем увеличивать число N. Если у нас получилось, что $\delta_N^2 \leq \varepsilon_n$ и $N > n$, нам следует уменьшить N.

Пример.

Среди аппроксимационных кривых, которые можно было провести через экспериментальные точки на рис.4.1, простейшей является прямая. Попробуем приблизить экспериментальные данные, измеренные в количестве $n+1$ штук, с помощью относительно простой функции $F(x)$

$$F(x) = C_1 + C_2 x; \quad \varphi_1(x) = 1; \quad \varphi_1(x) = x$$

Экспериментальные данные порождают столбцы

$$\varphi_1(x_p) = \left.\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}\right\} n+1, \quad \varphi_2(x_p) = \left.\begin{pmatrix} x_0 \\ \vdots \\ x_n \end{pmatrix}\right\} n+1$$

Матрица Грама нормальной системы уравнений

$$\begin{pmatrix} (\varphi_1,\varphi_1) & (\varphi_1,\varphi_2) \\ (\varphi_2,\varphi_1) & (\varphi_2,\varphi_2) \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} = \begin{pmatrix} (f,\varphi_1) \\ (f,\varphi_2) \end{pmatrix} \tag{21}$$

Запишем его по другому

$$\begin{pmatrix} \sum_{p=0}^{n} 1 & \sum_{p=0}^{n} x_p \\ \sum_{p=0}^{n} x_p & \sum_{p=0}^{n} x_p^2 \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} = \begin{pmatrix} \sum_{p=0}^{n} f(x_p)\cdot 1 \\ \sum_{p=0}^{n} f(x_p) x_p \end{pmatrix}$$

Если степень многочлена $F(x)$ равна n, то размерность матрица системы будет $n \times n$.

Пример.

Рассмотрим задачу среднеквадратичной аппроксимации методом НК.

2π—периодической функции (и ее периодическое продолжение) на равномерной сетке $\overline{\omega_{n-1}}$, покрывающей ее период $T = [0; 2\pi)$(рис.4.2). Точку 2π мы не учитываем, так как вследствие периодичности значение функции на ней совпадает со значением в точке 0.

Нужно с помощью системы тригонометрических функций $e^{ikx} = \varphi_k(x)$ получить функции в узлах сетки $x_p = \frac{2p}{n}p$. Эта система функций при весе $\rho = 0$ является ортогональной

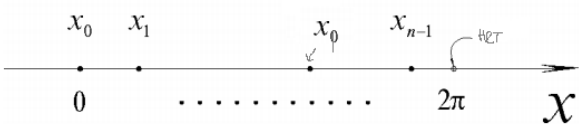

Рис. 4.2

$$(\varphi_k,\varphi_m)_{k\neq m} = \sum_{p=0}^{n-1} e^{ik\frac{2\pi}{n}p} e^{-im\frac{2\pi}{n}p} = \sum_{p=0}^{n-1} e^{i(k-m)\frac{2\pi}{n}p} =$$

$$= 1 + \underbrace{e^{i(k-m)\frac{2\pi}{n}}}_{q} + \ldots + \underbrace{e^{i(k-m)\frac{2\pi}{n}(n-1)}}_{q^{n-1}} = \frac{1 - e^{i(k-m)\frac{2\pi}{n}n}}{1 - e^{i(k-m)\frac{2\pi}{n}}} = 0$$

Квадрат нормы φ_k равен

$$(\varphi_k,\varphi_k) = n$$

Условия ортогональности позволяют переписать нашу нормальную систему (21) в виде

$$C_k(\varphi_k, \varphi_k) = (f, \varphi_k) = \sum_{p=0}^{n-1} f(x_p) e^{ik\frac{2\pi}{n}p}$$

Отсюда мы получаем, что представление функции $F_N(x)$ равно

$$F_N(x) = \sum_{k=1}^{N} C_k \varphi_k(x) = \sum_{k=1}^{N} \frac{f_k}{n} e^{ik\frac{2\pi}{n}x}$$

Сглаживание (фильтрация) экспериментальных таблиц методом наименьших квадратов

Наша цель состоит в том, чтобы, имея дискретное представление о некоторой информации, перестроить эти значения, получив другие значения в виде "сглаженных результатов". Пусть нам даны значения функции $F(x)$ в точках x_{k-1}, x_k, x_{k+1}, обозначим из f_{k-1}, f_k, f_{k+1} соответственно. Наша задача - перестроить данную нам таблицу данных так, чтобы новые значения функции $F(x)$ представляли собой значения функции $F_n(x)$. Количество параметров N, участвующих в таком среднеквадратичном приближении, удобно выбрать небольшим.

$$F(x) = C_1 + C_2(x - x_k)$$

Здесь присутствует три базисные функции: $\varphi_1(x) \equiv 1, \varphi_2(x) \equiv (x - x_k)$. Проецируясь на нашу сетку из трех узлов, эти функции порождают векторы размерности три. Эти векторы и участвуют в формировании матрицы нормальной системы и нахождения коэффициентов C_1, C_2. Не ограничивая общности предположим, что шаг сетки $h = const, \rho_k = 1$, запишем алгебраическую систему

$$\begin{pmatrix} (\varphi_1, \varphi_1) & (\varphi_1, \varphi_2) \\ (\varphi_2, \varphi_1) & (\varphi_2, \varphi_2) \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} = \begin{pmatrix} (f, \varphi_1) \\ (f, \varphi_2) \end{pmatrix}$$

$$\varphi_1(x_p) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad \varphi_2(x_p) = \begin{pmatrix} -h \\ 0 \\ h \end{pmatrix}, \quad (f, \varphi_1) = f_{k-1} + f_k + f_{k+1}, \quad (f, \varphi_1) = -hf_{k-1} + hf_{k+1}$$

$$\begin{pmatrix} 3 & 0 \\ 0 & 2h^2 \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} = \begin{pmatrix} f_{k-1} + f_k + f_{k+1} \\ -hf_{k-1} + hf_{k+1} \end{pmatrix}$$

$$F(x) = \frac{f_{k-1} + f_k + f_{k+1}}{3} + \frac{f_{k+1} - f_{k-1}}{2h}(x - x_k)$$

Тогда новое значение в точке: $F(x_k) = y_k = \frac{f_{k-1} + f_k + f_{k+1}}{3}$

Тем самым мы получили простейшую среднеквадратичную аппроксимацию.

5. Лекция 5. Вопросы численного дифференцирования и интегрирования функций.

Задача численного интегрирования функций заключается в вычислении приближенного значения определенного интеграла

$$I = \int_a^b f(x\rho(x)\,dx) \qquad \rho(x) > 0$$

Если найти первообразную функции $f(x)$ достаточно сложно, то задача вычисления интеграла состоит в замене подынтегральной функции $f(x)$ интерполирующей или аппроксимирующей функцией $g(x)$

$$\int_a^b g(x)\rho(x)\,dx + R[g]$$

Формулы численного вычисления однократного интеграла называются квадратурными формулами интерполяционного типа, где в качестве функции $g(x)$ выбирается интерполяционный многочлен.

Обозначим через $y_i = f(x_i)$ значение подынтегральной функции в различных точках $x_i \in \omega_n$ на $[a, b]$. В качестве приближенной функции $g(x)$ рассмотрим интерполяционный полином на ω_n в форме полинома Лагранжа:

$$g(x) = L_n(x) = \sum_{k=0}^{n} f_k l_k^{(n)}(x)$$

Получим формулу для приближенного значения интеграла

$$\tilde{I} = \int_a^b \left(\sum_{k=0}^{n} f(x_k) L_k^{(n)}(x) \right) \rho(x)\,dx = \sum_{k=0}^{n} C_k f(x_k)$$

x_k-узлы, C_k- веса

$$c_k = \int_a^b l_k^{(n)}(x)\rho(x)\,dx$$

Тогда интересующий нас интеграл I может быть вычислен с помощью квадратурной формулы

$$I = \tilde{I} + R[g]$$

Замечание. Узлы сетки нам заданы, поэтому базис Лагранжа определен однозначно, и базисные функции не зависят от значения интегрируемой функции $f(x)$. Это означает, что коэффициенты C_k вычисляются однократно, и дальше на этих сетках можно интегрировать все функции $f(x)$, которые доступны нам в постановке задачи.

Мы говорили, что погрешность полиномиальной интерполяции представляла собой следующую конструкцию

$$f(x) - P_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega(x) \quad \xi \in [a,b]$$

Тогда остаточный член

$$R[g] = \int_a^b \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega(x) = \frac{C(n)}{(n+1)!}f^{(n+1)}(\tilde{\xi})$$

Это позволяет утверждать, что для полиномов до порядка n включительно квадратурная формула точна, ведь в оценку остаточного члена входит $n+1$-ая производная. Наивысшая степень полинома, для которого квадратурная формула точна, называется степенью квадратурной формулы. Квадратурная формула интерполяционного типа имеет степень не ниже n. Вычисление весов $\{C_k\}$ квадратурной формулы можно проводить используя тот факт, что формула точна для $y = x^k$; $0 \le k \le n$. Получим СЛАУ для определения коэффициентов $\{C_k\}$

$$\int_a^b x^p\,dx = \sum_{k=0}^{n} C_k(x^p)\Big|_{x_k} = \frac{b^{p+1} - a^{p+1}}{p+1}, \quad p = \overline{1,n}$$

Определитель получившейся СЛАУ есть определитель Вандермонда системы $\{x^k\}$, посчитанный в узлах сетки ω_n

$$W(1, x, \dots, x^n) \neq 0$$

поэтому задача однозначно разрешима.

5.1. Квадратурные формулы Ньютона-Котесса

Рассматриваем равномерные сетки, отвечающие весу, равному единице

$$\rho(x) = 1; \quad \omega_n = \{x_n = x_0 + hn; \quad h = \frac{b-a}{n} \quad h = const\}$$

Посмотрим, как будет выглядеть базисный многочлен $l_k^{(n)}(x)$

$$l_k^{(n)}(x) = \frac{\omega(x)}{(x - x_k)\omega'(x_k)}$$

Где $\omega(x)$ - многочлен $n+1$-ой степени

$$\omega(x) = (x-x_0)(x-x_1)\dots(x-x_n) = h^{n+1}\underbrace{\left(\frac{x-x_0}{h}\right)}_{q}\left(\frac{x-x_1}{h}\right)\dots\left(\frac{x-x_n}{h}\right) =$$

$$= h^{n+1}q(q-1)(q-2)\dots(q-n)$$

Тогда производная от $\omega(x_k)$

$$\omega'(x_k) = \underbrace{(x_0 - x_k)\dots(x_{k-1} - x_k)}_{k} \cdot \underbrace{(x_{k+1} - x_k)\dots(x_n - x_k)}_{n-k} = h^n(-1)^k k!(n-k)!$$

В таком случае

$$C_k = \int_a^b l_k^{(n)}(x)\,dx = \frac{h^{(n+1)}}{h^{n+1}(-1)^k k!(n-k)!}\int_a^b \frac{q(q-1)(q-2)\dots(q-n)}{q-k}\,dx \Rightarrow$$

$$C_k = \frac{h(-1)^k}{k!(n-k)!}\int_0^n \frac{q(q-1)(q-2)\dots(q-n)}{q-k}\,dq \tag{22}$$

Формула (22) для вычисления весов квадратурной формулы Ньютона-Котесса. Обычно коэффициенты $\{C_k\}$, вычисленные таким образом, записывают через коэффициенты Котесса

$$C_k = (b-a)\mathcal{K}_k; \quad \mathcal{K}_p = \frac{(-1)^p}{p!(n-p)!n\int_0^n}\frac{q(q-1)(q-2)\dots(q-n)}{q-p}\,dq$$

А сама квадратурная формула Ньютона-Котесса (23) принимает вид

$$I = (b-a)\sum_{p=0}^{n}\mathcal{K}_p f(x_p) + R[g] \tag{23}$$

Для коэффициентов Котесса имеют место соотношения: 1) Если интегрируется функция тождественно равная единице, то квадратурная формула точна

$$\int_1^b 1\cdot\,dx = (b-a)\sum_{p=0}^{n}\mathcal{K}_p\cdot 1$$

Отсюда следует что

$$\sum_{p=0}^{n}\mathcal{K}_p = 1$$

2)

$$\mathcal{K}_p = \mathcal{K}_{n-p}$$

Попробуйте доказать самостоятельно.

5.2. Частные случаи формул Ньютона-Котесса

Квадратурная формула трапеций. Пусть интерполяционный полином Лагранжа $l_n(x)$ полином 1-ой степени (рис.5.1). Тогда

$$\begin{cases}\mathcal{K}_0 + \mathcal{K}_1 = 1\\ \mathcal{K}_0 = \mathcal{K}_1 = 1/2\end{cases}$$

Тогда квадратурная формула трапеции

$$\int_{x_0}^{x_1} f(x\,dx) = I_{\text{тр}} + R_{\text{тр}} = (x_1 - x_0)\left(\frac{1}{2}y_0 + \frac{1}{2}y_1\right) + R_{\text{тр}}$$

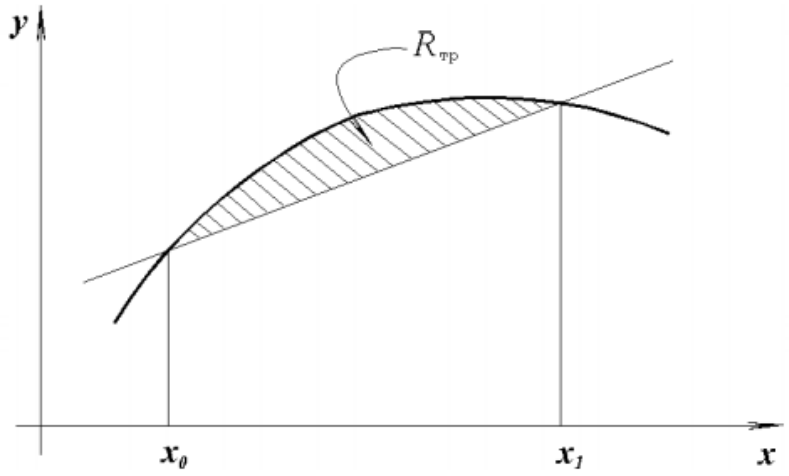

Рис. 5.1 – метод трапеций

При этом

$$I_{\text{тр}} = h\left(\frac{1}{2}f_0 + \frac{1}{2}f_1\right)$$

Оценим остаточный член в формуле трапеции

$$R_{\text{тр}} = \int_{x_0}^{x_1} \frac{f''(\xi)}{2!}(x - x_0)(x - x_1)\,dx = \frac{f''(\tilde{\xi})}{2!}\int_0^h t(t-h)\,dt = \frac{f''(\tilde{\xi})}{2}\left(\frac{h^3}{3} - \frac{h^3}{2}\right) =$$

$$= -\frac{h^3}{12}f''(\tilde{\xi}) \quad \xi \in (x_0, x_1)$$

Мы видим, что степень полученной формулы первая с остаточным членом $O(h)$, а значит она точна для полиномов до 1-го порядка включительно.

Квадратурная формула Симпсона (формула парабол)

Рассмотрим теперь случай когда $n = 2$. Сетка $\overline{\omega_2} = \{x_0, x_1, x_2\}$ содержит три узла(рис.5.2).

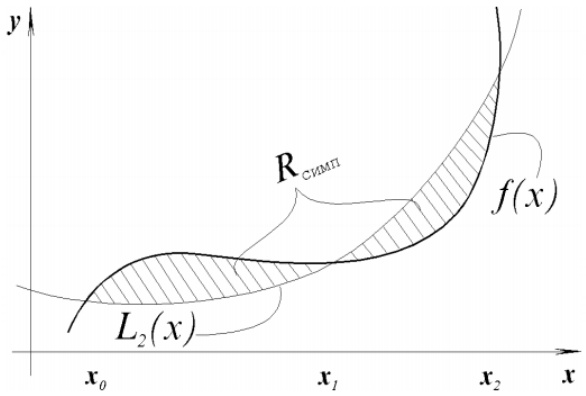

Рис. 5.2 – метод Симпсона

Соответствующая расчетная формула выглядит следующим образом:

$$\overline{I}_{\text{симп.}} = (x_2 - x_0) \sum_{p=0}^{2} \mathcal{K}_p f_p$$

У нас три коэффициента Котесса.

$$\mathcal{K}_0 + \mathcal{K}_1 + \mathcal{K}_2 = 1$$
$$\mathcal{K}_0 = \mathcal{K}_2$$

$$\mathcal{K}_2 = \frac{(-1)^2}{2! \cdot 0! \cdot 2} \int_0^2 \frac{q(q-1)(q-2)}{q-2}\, dq = \frac{1}{4}\left(\frac{q^3}{3} - \frac{q^2}{2}\right)\Bigg|_0^2 = \frac{1}{6}$$

$$\mathcal{K}_0 = \mathcal{K}_2 = \frac{1}{6}; \quad \mathcal{K}_1 = \frac{4}{6}$$

$$\overline{I}_{\text{симп.}} = 2h\left(\frac{1}{6}f_0 + \frac{4}{6}f_1 + \frac{1}{6}f_2\right) = \frac{h}{3}(f_0 + 4f_1 + f_2)$$

Для остаточного члена справедливо

$$R_{\text{симп.}} = \int_{x_0}^{x_2} \frac{f'''(\xi)}{\xi!}(x - x_0)(x - x_1)(x - x_2)\, dx \tag{24}$$

Если перейти к переменной $t = x - x_1$ и вынести $\frac{f'''(\xi)}{\xi!}$ за скобки по формуле среднего, то можно видеть, что выражение (24) равно нулю. Отсюда не следует, что погрешность по формуле Симпсона всегда равна нулю. Это означает, что в формуле нам нужны следующие слагаемые в разложении погрешности интерполяционного многочлена по производным функции $f(x)$. Тем самым в разложение есть производные более высоких порядков, а производные до порядка 3 включительно дают нулевой вклад. Значит формула точна и для многочленов 3-го порядка.

Получим остаточный член другим образом, представив его как функцию аргумента h, считая что интегрируемая функция обладает необходимым порядком гладкости

$$R_{\text{симп.}}(h) = \int_{x_1-h}^{x_1+h} f(x)\, dx - \frac{h}{3}(f(x_1 - h) + 4f(x_1) + f(x_1 + h)) \quad R(0) = 0$$

$$R'_{\text{симп.}}(h) = f(x_1 + h) + f(x_1 - h) - \frac{1}{3}(f(x_1 - h) + 4f(x_1) + f(x_1 + h)) -$$
$$-\frac{h}{3}(-f'(x_1 - h) + f'(x_1 + h))$$
$$R'(0) = 0$$

Далее нам нужно посчитать вторую и третью производную функции $R_{\text{симп.}}(h)$. Опустив соответствующие выкладки, запишем конечный результат

$$R_{\text{симп.}}(h) = -\frac{h^5}{90} y^{(4)}(\xi), \qquad (x_0, x_2)$$

Мы видим, что главный член асимптотики начинается с 4-ой производной, тем самым формула Симпсона имеет повышенную степень по отношению к интерполяционному многочлену.

Составные квадратурные формулы

Так как интеграл – аддитивная величина, вместо увеличения степени интерполяционного многочлена на отрезке $[a, b]$ нужно пытаться разбить область интегрирования на более мелкие отрезки, и на каждом из них выполнить интегрирование формулой относительно невысокого порядка. Для этого мы пользуемся составными квадратурными формулами, простейшая из которых – составная формула трапеций.

Общая формула трапеций

Разобьем отрезок $[a, b]$ на N частей длины $h = \frac{b-a}{N}$, совместив узлы интерполяции с узлами внешнего разбиения

$$I = \int_a^b f(x)\,dx = \sum_{k=1}^{N} \int_{x_{k-1}}^{x_k} f(x)\,dx =$$

$$= h\left(\frac{1}{2}y_0 + \frac{1}{2}y_1 + \frac{1}{2}y_1 + \frac{1}{2}y_2 + \ldots + \frac{1}{2}y_{N-1} + \frac{1}{2}y_N\right) + \sum_{k=1}^{N}\left(-\frac{h^3}{12}y''(\xi_k)\right) = \ldots$$

Эту формулу, как правило, упрощают, проводя вычисления во внутренних точках с коэффициентом 1, а во внешних с коэффициентом $\frac{1}{2}$.

$$= h\left(\frac{1}{2}y_0 + y_1 + \ldots + y_{N-1} + \frac{1}{2}y_N\right) + \sum_{k=1}^{N}\left(-\frac{h^3}{12}y''(\xi_k)\right)$$

Оценим выражение остаточного члена

$$m_2 = \min_{[a,b]} y''(x) \le y''(\xi_k) \le M_2 = \max_{[a,b]} y''(x)$$

$$m_2 \le \sum \frac{y''(\xi_k)}{N} \le M_2$$

$$\sum_{k=1}^{N} y''(\xi_k) = N y''(\xi)$$

$$R_{\text{тр.}} = -\frac{h^3}{12} N y''(\xi) = -\frac{b-a}{12} y''(\xi) h^2$$

Асимптотическая степень формулы равна двум, $R_{\text{тр.}} = O(h^2)$.

Составная формула Симпсона

Пусть $N = 2m$, т.е. на отрезке интегрирования находится $(2m+1)$ узел. Применим формулу Симпсона по каждому частичному сдвоенному промежутку длины $2h$:

$$[x_0, x_2],\ \ [x_2, x_4], \ldots,\ \ [x_{2m-2}, x_{2m}]$$

тогда

$$I = \int_a^b f(x)\,dx = \sum_{k=1}^{m} \int_{x_{2k-2}}^{x_{2k}} f(x)\,dx = \frac{h}{3}(y_0 + 4y_1 + y_2) + \frac{h}{3}(y_2 + 4y_3 + y_4) + \dots$$

$$\dots + \frac{h}{3}(y_{2m-2} + 4y_{2m-1} + y_{2m}) + \sum_{k=1}^{N} R_{trk}(x) = \frac{h}{3}(y_0 + y_{2m} + 4(y_1 + y_3 + \dots$$

$$\dots + y_{2m-1}) + 2(y_2 + y_4 + \dots + y_{2m-2})) + R_{sim}$$

Если $y \in C^{(4)}[a,b]$, то для оценки остаточного члена полученной формулы имеем

$$R_{sim} = -\frac{h^5}{90} \sum_{k=1}^{m} y^{(4)}(\xi_k), \quad \xi_k \in (x_{2k-2}, x_{2k})$$

Из аналогичных соображений получим, что $\exists \xi \in [a,b]$ и

$$y^{(4)}(\xi) = \frac{1}{m} \sum_{k=1}^{m} y^{(4)}(\xi_k),$$

что дает представление

$$R_{sim}(h) = -\frac{h^5}{90} m y^{(4)}(\xi) = -\frac{h^4}{180}(b-a) y^{(4)}(\xi),$$

что обеспечивает четвертый порядок точности и третью степень квадратурной формулы Симпсона.

6. Лекция 6.Квадратурные формулы Гаусса-Кристоффеля

Посмотрим на построение квадратурных формул интерполяционного типа несколько иначе

$$I = \int_a^b f(x)\rho(x)\,dx = \sum_{k=1}^{n} C_k f(x_k) + R[f] \tag{25}$$

Здесь удобно суммировать именно с $k = 1$ до n. Будем считать параметрами квадратурной формулы узлы $\{x_k\}$ и веса $\{C_k\}$ – в нашем распоряжении всего $2n$ параметров. Поставим вопрос о таком выборе параметра $\{x_k\}$, $\{C_k\}$, при котором квадратурная формула точна для многочленов максимально возможного порядка, по крайней мере, до $(2n-1)$ включительно. Покажем, как можно это сделать.

6.1. Выбор узлов квадратурной формулы

Мы знаем, что существуют системы классических ортогональных полиномов $\{P_n(x)\}$ с весом $\rho(x)$ на отрезке $[a,b]$

$$\int_a^b P_n(x)P_m(x)\rho(x)\,dx = \delta_{k,m}||P_k||^2$$

Корни $\{\mu_i\}$ полинома $P_k(x) = 0$ принадлежат отрезку $[a,b]$. Предположим, что система узлов $\{x_1, x_2, \ldots, x_n\}$, тогда мы можем рассмотреть многочлен степени n $\omega(x)$, корни которого совпадают с узлами

$$\omega(x) = (x - x_1)\ldots(x - x_n)$$

Тогда функция $\varphi(x) = \omega(x)P_m(x)$ при $m \le n-1$, многочлен степени не выше чем $2n-1$, поэтому формула (25) точна и остаточный член равен нулю

$$\int_a^b \varphi(x)\rho(x)\,dx = \int_a^b \omega(x)P_m(x)\rho(x) = \sum_{k=1}^{n} C_k \underbrace{\omega(x_k)}_{=0} P_m(x_k) = 0 \tag{26}$$

Таким образом, мы видим, что на интервале $[a,b]$ многочлен $\omega(x)$ ортогонален многочлену $P_m(x)$ для любого $m \le n-1$. Воспользуемся возможностью разложить произвольный многочлен по системе классических ортогональных многочленов

$$\omega(x) = \sum_{k=0}^{n} a(x)P_k(x) =$$

Продолжим формулу (26)

$$0 = \sum_{k=0}^{n} a_k \int_a^b P_k(x)P_m(x)\rho(x)\,dx = a_n||P_n||^2$$

Итак, все $a_m = 0$, кроме a_n и разложение $\omega(x)$ равны нулю. Разложение $\omega(x)$ имеет вид

$$\omega(x) = AP_n(x)$$

Мы получили возможность сформировать важный вывод. Узлы $\{x_i\}$ квадратурной формулы Гаусса-Кристоффеля нужно выбирать так, чтобы они совпадали с корнями ортогонального на $[a, b]$ с весом $\rho(x)$ многочлена $P_n(x)$.

$$P_n(\mu) = 0 \Leftrightarrow x_i = \mu_i \quad i = \overline{1, n}$$

6.2. Веса квадратурной формулы

Займемся вычислением индексов $\{C_k\}$ при известных узлах сетки. Рассмотрим базисный интерполяционный полином Лагранжа

$$l_k^{(n-1)}(x) = \frac{\omega(x)}{(x - x_k)\omega'(x_k)}$$

Это многочлен степени $n - 1$, тем самым формула Гаусса-Кристоффеля (25) точна

$$\int_a^b l_k^{(n-1)}(x)\rho(x)\,dx = \sum_{p=1}^{n} C_p \underbrace{l_k^{(n-1)}(x_p)}_{\delta_{p,k}} = C_k$$

Веса квадратурной формулы вычисляются по формулам весов интерполяционного типа. Тем самым мы показали, что формула (25) является формулой интерполяционного типа.

Существует теорема, доказывающая, что если интерполяционная сетка содержит n узлов и соответственно количество весовых коэффициентов равно n, то не удается построить квадратурную формулу, которая интегрировала произвольный многочлен порядка $n + 1$.

Также обратим внимание на то, что весовые коэффициенты в формуле Гаусса-Кристоффеля знакопостоянны. Чтобы убедиться в этом, посчитаем следующий интеграл

$$\int_a^b (l_k^{(n-1)}(x))^2\rho(x)\,dx = \sum_{p=1}^{n} C_p(l_k^{(n-1)}(x_p))^2 = C_k > 0$$

Интеграл слева больше нуля, значит и C_k больше нуля.

6.3. Простейший случай квадратурных формул Гаусса-Кристоффеля (формула средних прямоугольников)

Пусть мы используем сетку, содержащую всего один узел x_1, весовой коэффициент - C_1, весовая функция $\rho(x) \equiv 1$

$$\tilde{I} = \sum_{k=1}^{n} C_1 f(x_1)$$

Таким образом, формула должна быть точна для многочленов до первого порядка включительно. Выполним интегрирование и решим систему

$$\begin{cases} \int_a^b 1\,dx = b-a \\ \int_a^b x\,dx = \frac{b^2-a^2}{2} = C_1 x_1 \end{cases} \Rightarrow x_1 = \frac{b+a}{2}$$

$$\int_a^b f(x)\,dx = (b-a) f\left(\frac{a+b}{2}\right) + R_{\text{пр.}}$$

Тем самым мы получили формулу средних прямоугольников. Сопроводим рис.6.1 наше интегрирование.

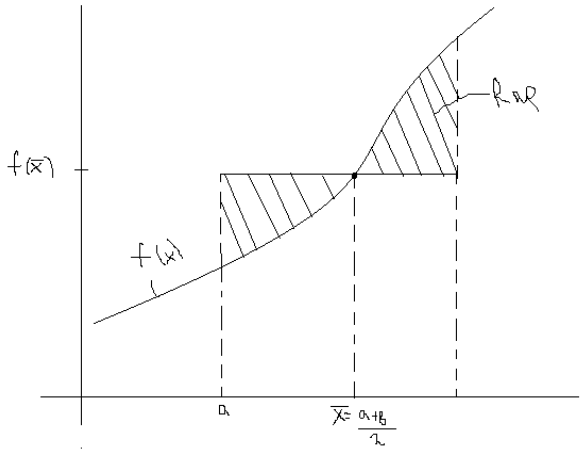

Рис. 6.1 – метод средних прямоугольников

Рассмотрим, что представляет собой остаточный член $R[h]$ в зависимости от длины отрезка интегрирования

$$R[h] = \int_{\overline{x}-\frac{h}{2}}^{\overline{x}+\frac{h}{2}} f(t)\,dt - hf(\overline{x})$$

Из этого сразу делаем вывод, что $R(0) = 0$, теперь вычислим производные

$$R'(h) = \left(F\left(\overline{x}+\frac{h}{2}\right) - F\left(\overline{x}-\frac{h}{2}\right)\right)' - f(\overline{x}) = f\left(\overline{x}+\frac{h}{2}\right)\frac{1}{2} + f\left(\overline{x}-\frac{h}{2}\right)\frac{1}{2} - f(\overline{x})$$

$R'(0) = 0$, вычислим вторую производную

$$R''(h) = \frac{1}{4}f'\left(\overline{x}+\frac{h}{2}\right) - \frac{1}{4}f'\left(\overline{x}-\frac{h}{2}\right) = \frac{1}{4}f''(\xi)h$$

Тогда

$$R'(h) = \int_0^h R''(t)\,dt = \frac{1}{4}f''(\xi)\frac{h^2}{2}$$

$$R(h) = \int_0^h R'(t)\,dt = \frac{f''(\xi)}{24}h^3$$

Окончательно имеем

$$\int_a^b f(x)\,dx = (b-a)f\left(\frac{a+b}{2}\right) + \frac{f''(\xi)}{24}h^3$$

В качестве дополнительного задания попробуйте получить формулу составных средних прямоугольников.

6.4. Апостериорная оценка погрешности

Предположим, что мы занимаемся вычислением величины $z(x)$ пользуясь расчетом $\xi(x,h)$ характеризующегося параметром h. Так как наши расчеты имеют приближенный характер, у нас добавляется главный член асимптотики и члены более высоких порядков.

$$z(x) = \xi(x,h) + \varphi(x)h^p + O(h^{p+1}) = \xi(x,h) + O(h^p)$$

Для интересующей нас величины выполним расчет, но сетку (или соответствующий параметр) изменим на q

$$x(x) = \xi(x,qh) + \varphi(x)(qh)^p + O(h^{p+1})$$

Теперь мы можем исключить главный член асимптотики с тем же порядком точности

$$\varphi(x)h^p(q^p-1) = \xi(x,h) - \xi(x,qh) + O(h^{p+1})$$

Отсюда мы получаем первую формулу Рунге

$$\varphi(x)h^p = \frac{\xi(x,h) - \xi(x,qh)}{q^p-1} + O(h^{p+1}) \tag{27}$$

Теперь мы можем рассчитать $z(x)$ с более высоким порядком точности

$$z(x) = \xi(x,h) + \frac{\xi(x,h) - \xi(x,qh)}{q^p-1} + O(h^{p+1}) \tag{28}$$

Важно то, что формула Рунге применима только тогда, когда мы знаем структуру главного члена асимптотики.

Пример. Уточним порядок точности вычисления интеграла по формуле Симпсона.

$$I = \int_a^b f(x)\,dx = I_h + \left(-\frac{f^{(4)}(\xi)}{180}\right)h^4 = I_h + C\cdot h^4 + O(h^5) =$$

Главный член асимптотики – величина порядка h^4. Если выполнить расчет этой же величины при $q=2$, то есть увеличив шаг сетки в два раза I_{2h}, и воспользоваться формулой (28)

$$= I_h + \frac{I_h - I_{2h}}{2^4-1} + O(h^5)$$

Используя два расчета, мы уточнили формулу вычисления интеграла о формуле Симпсона.

Известно, что весу $\rho(x) \equiv 1$ на отрезке $[a, b]$ отвечает система классических ортогональных многочленов Лежандра $L_n(x)$. Мы получим эту систему методом ортогонализации стандартного набора мономиальных функций x^k. Первое, что нужно, – это сделать замену переменных величины x, меняющуюся на отрезке $[a, b]$, на величину t, меняющуюся на отрезке $[-1, 1]$

$$\int_a^b f(x)\,dx = \left| \begin{array}{cc} x = ct + d & dx = cdt \\ a = -c + d & c = \frac{b-a}{2} \\ b = c + d & d = \frac{b+a}{2} \\ x = \frac{b-a}{2}t + \frac{b+a}{2} & t \in [-1, 1] \end{array} \right| =$$

$$\frac{b-a}{2} \int_{-1}^{1} F(t)\,dt = \frac{b-a}{2} C_1 F(t_1) + R_{\text{Пр.}}$$

Коэффициент C_1 нужно вычислять, интегрируя базисный многочлен $l^{(0)}(t)$.

Посмотрим, что из себя представляют многочлены Лежандра нулевого и первого порядка

$$P_0(t) = 1; \qquad \int_{-1}^{1} t\,dt = 0 \Rightarrow t \perp 1 \Rightarrow P_1(t) = t$$

Корень многочлена $P_1(t)$ равен $\mu_1 = 0$, тем самым узел t_1 окажется равным нулю. Осталось посчитать, чему равен вес C_1.

$$C_1 = \int_{-1}^{1} l^{(0)}(t)\,dt = \int_{-1}^{1} dt = 2$$

Окончательно имеем

$$\int_a^b f(x)\,dx = (b-a) f\left(\frac{a+b}{2}\right) + R_{\text{Пр.}}$$

6.5. Численное дифференцирование

Немного обсудим вопросы численного дифференцирования. Пусть стоит задача заменить вычисление производных функции $f(x)$ вычислением производных приближающей ее функции $g(x)$

$$f^{(k)} \to g^{(k)}$$

В формулах интерполяционного типа в качестве приближения $g(x)$ мы использовали интерполяционный многочлен

$$g(x) = L_n(x) + N_n(x) = P_n(x)$$

Таким образом, задача дифференцирования функции $g(x)$ в такой форме достаточно тривиальна. Погрешность интерполяционной формулы

$$|f(x) - P_n(x)| = \left| \frac{f^{(n+1)}(\xi)}{(n+1)!} \omega(x) \right| = O(h^{n+1})$$

Опираясь на эту формулу, можно посмотреть, что из себя представляет погрешность при численном дифференцировании

$$|f^{(k)}(x) - P_n^{(k)}(x)| \leq \left| \frac{f^{(n+1)}(\xi)}{(n+1)!} C_{n+1}^k k! (\max|x - x_i|)^{n+1-k} \right| = O(h^{n+1-k})$$

В качестве приближающей функции удобно использовать сплайн $S_3^1(x) \equiv S(x)$. Была сформулирована теорема об оценке как самого сплайна, так и его производных в равномерной метрике на отрезке $[a, b]$.

$$||f(x) - S(x)||_{[a,b]} \leq C_0 M_4 h^4$$
$$||f'(x) - S'(x)||_{[a,b]} \leq C_1 M_4 h^3$$
$$||f''(x) - S''(x)||_{[a,b]} \leq C_2 M_4 h^2$$

7. Лекция 7.Решение нелинейных уравнений

Одной из наиболее распространенных вычислительных задач является задача нахождения корня уравнения.

Ограничимся рассмотрением двух случаев.

1) Пусть $x \in R_1$. Тогда мы решаем уравнение

$$f(x) = 0 \tag{29}$$

2) Пусть $x \in R_m$, тогда мы рассматриваем систему, где каждое из уравнений зависит от $x_1, x_2, \ldots, x_m$

$$f_k(x_1, x_2, \ldots, x_m), \quad k = \overline{1, m} \tag{30}$$

7.1. Метод деления отрезка пополам

Начнем с описания постановки задачи (29), ее приближенное решение можно получить методом деления отрезка пополам (метод вилки), для этого сформулируем теорему.

Теорема 7 Пусть $f(x) \in C[a, b]$, тогда для $\forall y$ принадлежащему $\langle f(a), f(b)\rangle$ найдется хотя бы одна такая точка $x \in [a, b]$, что

$$f(x) = y$$

замечание. Угловые скобки означают, что мы не знаем точно, какой из концов отрезка $f(a), f(b)$ расположен левее или правее.

Требования теоремы можно усилить взяв y принадлежащее $[\min_{[a,b]} f(x), \max_{[a,b]} f(x)]$.

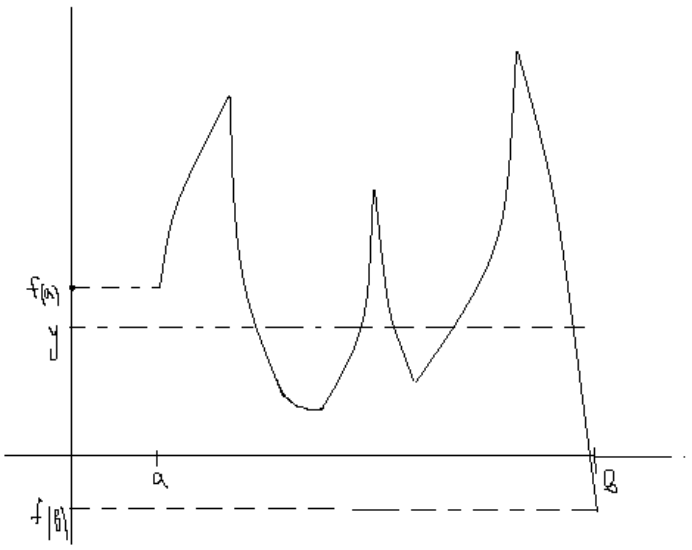

Рис. 7.1 – иллюстрация к теореме 7

На применении теоремы 5 основан метод вилки для поиска корней уравнения $f(x) = 0$. Локализуют интервал, на концах которого функция принимает значения разных знаков $f(x_{\text{лев}})f(x_{\text{пр}}) < 0$. Если сразу оказалось, что $x_{\text{лев}}$ или $x_{\text{пр}}$, дает нам ноль, то корень найден. Если нет, то из теоремы следует, что существует точка в которой $f(x) = 0$. Вычисляем середину нашего исходного интервала $[x_{\text{лев}}, x_{\text{пр}}]$ и ищем значение $f(x)$ в этой точке, $f\left(\frac{x_{\text{лев}}+x_{\text{пр}}}{2}\right) = f(x_{\text{ср}})$. Если $f(x_{\text{ср}}) = 0$, то корень найден. Если нет, то проверяем на каком интервале знак функции поменялся, то есть какое из неравенств

$$f(x_{\text{ср}})f(x_{\text{лев}})$$

$$f(x_{\text{ср}})f(x_{\text{пр}})$$

меньше нуля. Выбираем интервал, на котором происходит смена знака и повторяем процедуру, описанную выше. Тогда если на нулевом шаге корень x^* принадлежал исходному интервалу $[a, b]$, то на k-ом шаге этот корень принадлежит интервалу, длина которого $L_k = \frac{b-a}{2^k}$. Таким образом, выбирая в качестве приближенного значения корня середину отрезка, можно построить последовательность x^k (середина отрезка на k-ом шаге) которая будет сходиться к x^*, в силу теоремы о стягивающейся системы сегментов. Погрешность соответствующего приближения, по крайней мере, не превосходит длины L_k, тем самым мы можем с заданной точностью ε решить задачу приближения корня x^*.

7.2. Метод последовательного приближения. Теорема о непрерывном сжатии

В большинстве случаев мы будем стараться свести задачу (29) к задаче о неподвижной точке отображения $\varphi(x)$

$$f(x) \Leftrightarrow \varphi(x) \tag{31}$$

Эти задачи равносильны, то есть имеют одинаковые корни. Для решения задачи (31) можно предложить метод последовательных приближений (МПП)

$$x_{k+1} = \varphi(k), \quad k = \overline{1, \infty} \tag{32}$$
$$x_0 = x^{(0)}$$

Нас интересует случай, когда последовательность (32) сходится к неподвижной точке, которая удовлетворяет уравнению (31)

Теорема 8. Принцип сжимающего отображения

Пусть φ непрерывное сжатие полного метрического пространства M, то существует единственная неподвижная точка $x^* \in M$ и она является пределом последовательности:

$$x_{k+1} = \varphi(x_k), \quad k = \overline{1, \infty}$$
$$\lim_{k\to\infty} x_n = x^*$$

Доказательство:

Пусть была построена последовательность $\{x_k\}$ по методу последовательных приближений. Выбрав произвольную точку x_0, покажем, что эта последовательность фундаментальна. Пусть n, m произвольные натуральные числа $(m > n)$, тогда

$$\rho(x_m, x_n) = \rho(\varphi(x_{m-1}), \varphi(x_{n-1})) \le \rho(x_{m-1}, x_{n-1}) \le q^2\rho(x_{m-2}, x_{n-2}) \le \dots$$
$$\dots \le q^n\rho(x_{m-n}, x_0)$$

Расстояние между точками $\rho(x_{m-n}, x_0)$ по неравенству треугольника меньше или равно длине ломаной соединяющей эти точки, поэтому справедлива оценка

$$\rho(x_0, x_{m-n}) \le \rho(x_0, x_1) + \rho(x_1, x_2) + \dots + \rho(x_{m-n-1}, x_{m-n}) \le$$

$$\leq \rho(x_0,x_1)+q\rho(x_0,x_1)+q^2\rho(x_0,x_1)+\ldots+q^{m-n-1}\rho(x_0,x_1)\leq \rho(x_0,x_1)\frac{1}{1-q}$$

Тогда

$$\rho(x_m,x_n)\leq \frac{q^n}{1-q}\rho(x_0,x_1)<\varepsilon,\quad \forall m,\forall n>N(\varepsilon)$$

получив такую оценку, мы можем утверждать, что последовательность фундаментальна, и в силу полноты пространства M нашей последовательности

$$\exists x^*;\quad \lim_{k\to\infty}x_k=x^*$$

Покажем, что эта точка x^* является неподвижной точкой отображения φ

$$x^*=\lim_{k\to\infty}x_k=\lim_{k\to\infty}\varphi(x_{k-1})=\varphi\left(\lim_{k\to\infty}x_{k-1}\right)=\varphi(x^*)$$

Теперь докажем единственность неподвижной точки. Для этого предположим, что существует еще одна неподвижная точка $\overline{x^*}$, тогда

$$\rho(x^*,\overline{x^*})=\rho(\varphi(x^*),\varphi(\overline{x^*}))\leq\rho(x^*,\overline{x^*})$$

Так как $\rho(x^*,\overline{x^*})>0$, на него можно сократить, и тогда получим

$$q\geq 1$$

А значит это отображение не является сжатием. Получаем противоречие, которое доказывает единственность неподвижной точки.

Напишем некоторые характеристики полученной сходимости:

1)

$$\rho(x_n,x^*)\leq\frac{q^n}{1-q}\rho(x_0,x_1)$$

2)

$$\rho(x_n,x^*)=\rho(\varphi(x_{n-1}),\varphi(x^*))\leq q\rho(x_{n-1},x^*)\leq\ldots\leq q^n\rho(x_0,x^*)$$

Будем говорить, что последовательность обладает сходимостью порядка p, если

$$\rho(x_{n+1},x^*)=O([\rho(x_n,x^*)]^p)$$

Тогда в нашем случае

$$\rho(x_n,x^*)=O(\rho(x_{n-1},x^*))$$

Поэтому построенная по методу МПП итерационная последовательность имеет степень не ниже первого порядка точности.

3)

$$\rho(x_n,x^*)=\rho(\varphi(x_{n-1}),\varphi(x^*))=\rho(\varphi(x_{n-1}),x^*)\leq\rho(\varphi(x_{n-1}),\varphi(x_n))+\rho(\varphi(x_n),x^*)\leq$$

$$\leq q\rho(x_{n-1},x_n)+q\rho(x_n,x^k)$$

Из этого получаем оценку

$$\rho(x_n, x*) \leq \frac{q}{1-q}\rho(x_{n-1}, x_n)$$

Часто считают, что при

$$\frac{q}{1-q}\rho(x_{n-1}, x_n) < \varepsilon$$

приближение достаточно точно, и точка x^* достигнута. Это не всегда является верным, например, если процесс обладает медленной сходимостью ($q \approx 1$).

Пример. Метод релаксации

Решаем уравнение $f(x) = 0$. Пускай на каком-то шаге последовательного приближения мы получили x_n и $f(x_n) \neq 0$. Вычисленная величина $f(x_n)$, в которой функция не обращается в ноль, называется невязкой уравнения. Выберем следующее приближение x_{n+1} таким образом, чтобы компенсировать невязку

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

Тогда преобразуем уравнение $f(x) = 0$

$$\tau(x)f(x) = x - x \Rightarrow x = x + \tau(x)f(x) \equiv \varphi(x)$$

Тем самым мы перешли к уравнению, характеризующемся как задача от неподвижной точки отображения φ. Процесс, который мы описали, носит названия метода релаксации

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

Здесь итерационный параметр $h = const$. Суть метода состоит в том, чтобы релаксировать невязку предыдущего приближения.

Пример. Метод Ньютона или метод касательных

Вместо того, чтобы решать исходное уравнение $f(x) = 0$, попытаемся заменить это уравнение, находясь в окрестности очередного шага x_n итерационного процесса

$$f(x) = f(x_n + \triangle x) = f(x_n) + f'(x_n) \triangle x + O(|\triangle x|)$$

Пренебрегая членами более высокого порядка малости, будем решать уравнение

$$y(x) = f(x_n) + f'(x_n)(x - x_n) = 0,$$

и корень этого уравнения брать как точку x_{n+1}. Нетрудно видеть, что это уравнение касательной к функции $f(x)$ в точке x_n (рис. 7.2 а)).

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

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

Пример. Метод секущих

Этот метод получают из метода Ньютона заменой $f(x)$ в окрестности точки x_n интерполяционным многочленом, приближающим нашу функцию (рис. 7.2 б)).

$$N_1(x) = f(x_n) + (x - x_n)f(x_n, x_{n-1})$$

Делаем замену

$$f(x) = 0 \to N_1(x) = 0$$

Корень нового уравнения x_{n+1}

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

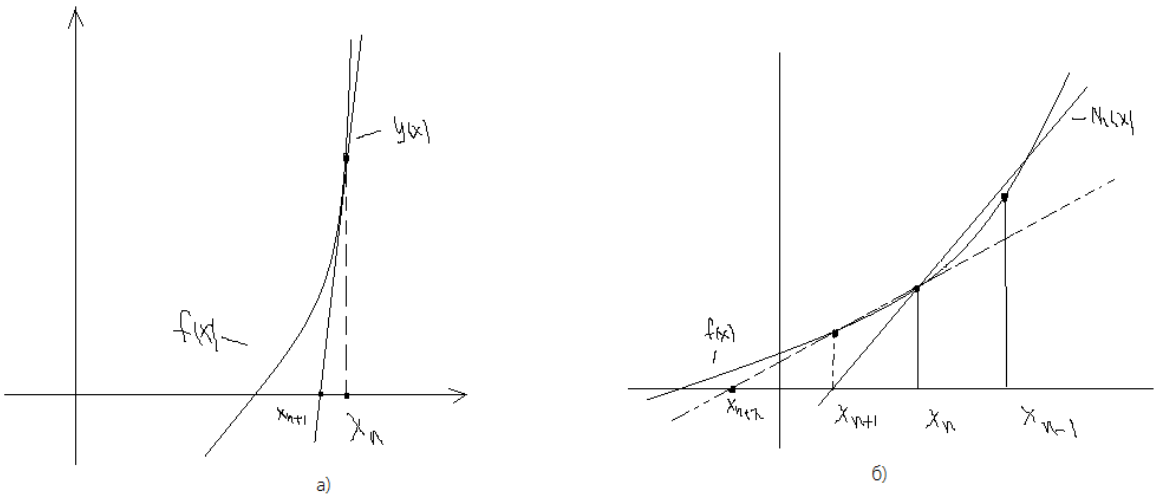

Рис. 7.2 – примеры поиска x_{n+1} разными способами

Пример. Метод парабол

Метод заключается в том, чтобы уравнение $f(x) = 0$ заменить на уравнение $N_2(x) = 0$

$$f(x_n) + (x - x_n)f(x_n, x_{n-1}) + (x - x_n)(x - x_{n-1})f(x_n, x_{n-1}, x_{n-2}) = 0$$

Это квадратное уравнение относительно величины $z = x - x_n$. Пусть $z_{\min}$ – минимальный из корней уравнения, тогда

$$x_{n+1} = x_n + z_{\min}$$

Задача. Записать, как будет выглядеть минимальный из корней параболы, несмотря на знак дискриминанта.

Пример. Метод обратной интерполяции

Метод обратной интерполяции использует интерполяцию, обратную к $f(x)$ функции $x = g(y)$

$$x^* : \quad f(x^*) = 0 \Leftrightarrow g(0) = x^*$$

Ставится задача вычисления значения $g(y)$ в нуле: $x_{n+1} = g(0)$. Пусть известны приближенные значения $x_0, x_1, \ldots, x_n$, соответствующие значениям функции

$$y_0 = f(x_0), y_1 = f(x_1), \ldots, y_n = f(x_n)$$

Построим на сетке $\{y_n\}_{n=\overline{1,n}}$ интерполяционный многочлен Лагранжа $L_n(y)$ для обратной функции $g(y)$

$$g(y) \approx L_n(y) = \sum_{k=1}^{n} x_k \frac{\omega(y)}{(y - y_k)\omega'(y_k)},$$

тогда

$$x_{n+1} = L_n(0)$$

Задача. Для случая $n = 1$ (т.е. двух точек $x_n, x_{n-1} \to g(y) = L_1(y)$) построить x_{n+1} (это метод секущих). На практике обычно уравнение имеет не единственный корень. Тогда имеет место рассмотрение отображения не на всем множестве M. Вопрос о локализации корня решает следующая теорема:

Теорема 9

Пусть Ω - открытая область полного метрического пространства M, и пусть на Ω задано сжимающее отображение $\varphi : \Omega \Rightarrow M$. Тогда для существования неподвижной точки x^* отображения φ в области Ω, $x^* \in \Omega$ необходимо и достаточно, чтобы нашелся в области Ω замкнутый шар $\overline{B}(x_0, r) \subset \Omega$, для которого

$$\rho(x_0, \varphi(x_0)) \leq (1 - q)r \tag{33}$$

Доказательство

Необходимость. Пусть в Ω существует неподвижная точка x^* для отображения φ. Поскольку Ω – открытая область, x^* - внутренняя точка области Ω, существует шар $\overline{B}(x^*, r) \subset \Omega$, тогда

$$\rho(x^*, \varphi(x^*)) = 0 \leq (1 - q)r$$

и подавно для $x_0 = x^*$

Достаточность. Рассмотрим произвольную точку $x \in \overline{B}(x_0, r)$, для которой выполнены требования (33)

$$\rho(x_0, \varphi(x)) \leq \rho(x_0, \varphi(x_0)) + \rho(\varphi(x_0), \varphi(x)) \leq (1 - q)r + qr = r$$

Таким образом, для любой точки, находящейся внутри нашего шара, сжатие φ переводит эту точку в шар. То есть если отображение φ рассматривать только на области $\overline{B}(x_0, r)$, то мы можем вернуться к теореме 1. Поскольку $\overline{B}(x_0, r)$ является замкнутым полным метрическим подпространством, тем самым наше непрерывное отображение имеет только одну неподвижную точку, и при любом начальном значении расположенным в этом шаре метод последовательных приближений сходится к корню.

В дальнейшем мы будем отождествлять $\overline{B}(x_0, r)$ с отрезком $[a, b]$. Понятно, что его центром является точка $x_0 = \frac{a+b}{2}$, а радиусом величина $r = \frac{b-a}{2}$. Перепишем условия (33) для отрезка $[a, b]$

$$|x_0 - \varphi(x_0)| \leq (1 - q)\frac{b - a}{2}$$

Для формулировки *достаточного* условия сжимающего отображения воспользуемся формулой Лагранжа конечных приращений

$$|\varphi(x_1) + \varphi(x_2)| = |\varphi'(\xi)(x_1 - x_2)| \leq |\max \varphi'| \cdot |x_1 - x_2| \leq M_1 |x_1 - x_2|$$

Итак, отображение φ является сжимающим, если

$$|\varphi'(x)| \leq q < 1$$

Чаще всего пользуются достаточным условием следующего вида. Пусть $\varphi \in C_1[a, b]$, если

$$|\varphi'(x^*)| < 1$$

тогда мы можем утверждать, в силу сохранения знака непрерывной функции, что найдется такое ε что

$$|x - x^*| < \varepsilon \to |\varphi'(x)| \leq q < 1$$

Сформулируем достаточные условия для метода релаксаций

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

$$\varphi'(x) = 1 + \tau f'(x)$$

$$|\varphi'(x^*)| = |1 + \tau f'(x^*)| < 1$$

$$-1 < 1 + \tau f'(x^*) < 1$$

Окончательно преобразуя, получим

$$-2 < \tau f'(x^*) < 0 \tag{34}$$

Как правило, это условие интерпретируют как условие на выбор параметра τ. Если мы имеем оценку для модуля производной, тогда мы можем найти итерационный параметр τ так, чтобы удовлетворить условию (34), что гарантирует локализацию и сходимость метода релаксации к корню x^*, если мы находимся в той области пространства, где выполнено неравенство (34) и для произвольной точки x.

Сформулируем достаточные условия для метода Ньютона

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

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

$$\varphi' = 1 - \frac{f'f' - ff''}{f'^2} = \frac{ff''}{f'^2}\bigg|_{x=x^*} = 0$$

$$x_{n+1} - x^* = \varphi(x_n) - \varphi(x^*) = \varphi'(x^*) \triangle x + \ldots + \frac{\varphi^{(p-1)}(x^*)}{(p-1)!}(x_n - x^*)^{p-1} + \frac{\varphi^{(p)}(\xi)}{p!}(x_n - x^*)^p$$

Если итерационная функция $\varphi(x)$ обладает свойством

$$\varphi(x^*) = \varphi'(x^*) = \varphi''(x^*) = \ldots = \varphi^{(p-1)}(x^*) = 0$$

тогда

$$|x_{n+1} - x^*| = O(|x_n - x^*|^p)$$

Тем самым метод Ньютона – не ниже второго порядка сходимости. Квадратичная сходимость обеспечивается в некоторой окрестности корня.

7.3. Итерационные методы решения систем нелинейных уравнений

Обобщим рассмотренные ранее случаи для задачи (30). Напомним, как выглядит уравнение

$$f_k(x_1, x_2, \ldots, x_m), \quad k = \overline{1, m} \Leftrightarrow F(x) = 0$$

Каноническая форма записи одношагового итерационного метода такова:

$$\underset{m\times m}{A_{k+1}} \frac{\overset{(k+1)}{\vec{x}} - \overset{(k)}{\vec{x}}}{\tau_{k+1}} - F\left(\overset{(*)}{\vec{x}}\right) = 0$$

Где A_{k+1} итерационная невырожденая матрица размерности $m \times m$.Подстановка очередной итерации в $F(x)$ порождает невязку. К следующей итерации мы хотим перейти так, чтобы компенсировать взвешенную с помощью τ невязку.

Ограничимся простейшим случаем, когда одношаговый итерационный метод является стационарным.

метод релаксации

$$\overset{(k+1)}{\vec{x}} = \overset{(k)}{\vec{x}} + \tau F\left(\overset{(*)}{\vec{x}}\right)$$

То есть это задача о неподвижной точки

$$x = \Phi(x) = Ex + \tau F(x)$$

Все величины являются векторными.

метод Ньютона

Уравнение $F(x) = 0$ в окрестности точки $\overset{(k)}{\vec{x}}$ (k-ая итерация это вектор) заменяют формулой Тейлора

$$F(x) = F\left(\overset{(*)}{\vec{x}}\right) + dF\left(\overset{(*)}{\vec{x}}\right) + \frac{1}{2!}d^2F(\vec{\xi})$$

Пренебрегая слагаемыми выше первого порядка мы получаем приблизительное уравнение относительно приращения аргумента $\overset{(k+1)}{\vec{x}} = \overset{(k)}{\vec{x}} + \vec{\triangle x}$

$$F_j\left(\overset{(k)}{\vec{x_1}} + \ldots + \overset{(k)}{\vec{x_m}}\right) + \sum_{p=1}^{n} \frac{\partial F_j}{\partial x_p}\left(\overset{(k)}{\vec{x}}\right)\left(\overset{(k+1)}{\vec{x_p}} - \overset{(k)}{\vec{x_p}}\right) = 0$$

Таким образом, мы получаем СЛАУ с невырожденной матрицей J

$$\frac{dF}{dx} = J = \frac{D(F_1, \ldots, F_m)}{x_1, \ldots, x_m} \neq 0$$

$$\vec{x}^{(k+1)} = \vec{x}^{(k)} - \left(\frac{dF}{dx}\right)^{-1} F\left(\vec{x}^{(k)}\right) = 0$$

Напомним, что в случае одного переменного мы получили

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

Единственное отличие в этом случае, в том что мы не можем выполнять операцию деления, поэтому нужно искать обратную матрицу.

Сформулируем теперь достаточные условия. Пускай у нас есть отображение $\vec{x} = \Phi(\vec{x})$(задача о неподвижной точке), оценим когда это отображение будет является сжатием

$$||\Phi(\overline{x}) - \Phi(\overline{\overline{x}})|| = \left|\left|\frac{d\Phi}{dx}(\xi)(\overline{x} - \overline{\overline{x}})\right|\right| \leq \underbrace{\left|\left|\frac{d\Phi}{dx}(\xi)\right|\right|}_{\Phi'(x)} \cdot \left|\left|(\overline{x} - \overline{\overline{x}})\right|\right|$$

Если

$$||\Phi(x)|| \leq q < 1$$

тогда отображение является сжимающим

Рассмотрим оценку для итерационной функции в методе Ньютона в случае многих переменных.

$$\Phi = Ex - (F'(x))^{-1} F(x)$$

$$\Phi' = E - \left\{\frac{d}{dx}(F')^{-1}F + \underbrace{(F')^{-1}\frac{dF}{dx}}_{=E}\right\} = -\frac{d}{dx}(F')^{-1}F\Big|_{x*} = 0$$

Таким образом, метод Ньютона обладает не ниже чем вторым порядком сходимости.

Получим оценки также для метода релаксации.

$$\vec{x}^{(k+1)} = \vec{x}^{(k)} + \tau F\left(\vec{x}^{(k)}\right)$$

$$\Phi = Ex + \tau F(x)$$

$$\Phi' = E + \tau F'(x)$$

$$||\varphi'(x^*)|| < 1$$

$$\Big|\,||E|| - ||\tau F'||\,\Big| \leq ||E + \tau F'(x^*)|| < 1$$

Окончательно получаем

$$-2 < -\left|\tau\right|||F'(x^*)|| < 0$$

Часто возникает проблема нахождения связи между погрешностью по аргументу и по значению интересующей нас функции.

$$0 \approx F\left(\overset{(n)}{x}\right) - F(x^*) = F'(\xi)(\overset{(n)}{x} - x^*)$$

$$||F(\overset{(n)}{x})|| \leq M_1||\overset{(n)}{x} - x^*||, \quad M_1 = \max_{\overline{B}}||F'(x)||$$

$$x^n - x^* = (F'(\xi))^{-1} \cdot F(\overset{(n)}{x})$$

$$||x^n - x^*|| \leq N_1 = ||F(\overset{(n)}{x})||; \quad N_1 \max_{\overline{B}}||(F'(x))^{-1}||$$

Мы получили связь между величиной ошибки $||x^n - x^*||$ и величиной невязки $||F(\overset{(n)}{x})||$ и наоборот.

8. Лекция 8. Основные методы решения уравнений. Метода последовательного исключения Гаусса.

Тема лекции основные задачи линейной алгебры. Ограничимся рассмотрением следующей постановки задачи

$$Ax = f; \quad A : R^n \Rightarrow R^N \tag{35}$$

Посмотрим, какими методами мы можем воспользоваться для ее решения. Как правило, в эту же задачу мы включаем вопрос о построении определителя матрицы A, поскольку теорема о существовании и единственности решения связана с его вычислением

$$detA = \triangle \neq 0$$

Формально, в случае невырожденного определителя мы можем построить решение задачи (35) либо с помощью обратной матрицы

$$x = A^{-1}f \Rightarrow$$

либо с помощью формул Крамера

$$(x)_k = \frac{\triangle_k}{\triangle}$$

где $\triangle_k$ определитель матрицы A в которой k-ый столбец заменен на столбец свободных членов.

Также формально мы можем ответить на вопрос об устойчивости задачи (35), варьируя переменную x получим

$$\delta A \cdot x + A \cdot \delta x = \delta f$$

Тем самым получим формальное представление влияния соответствующих погрешностей на погрешность результата

$$\delta x = A^{-1}(\delta f - \delta A \cdot x)$$

Из этого мы можем сделать вывод

$$||\delta x|| \to 0 \quad \text{при} \quad ||\delta A||, ||\delta f|| \to 0$$

Предположим, что матрица A известна нам точно $\delta A = 0$ и погрешность решения связана лишь с величиной δf, в таком случае

$$\delta x = A^{-1}\delta f$$

$$||\delta x|| = ||A^{-1} \cdot \delta f|| \leq ||A^{-1}|| \cdot ||\delta f||$$
$$||f|| = ||A \cdot x|| \leq ||A|| \cdot ||x||$$

$$||\delta x|| \cdot ||f|| \leq ||A|| \cdot ||A^{-1}|| \cdot ||\delta f|| \cdot ||x||$$
$$\frac{||\delta x||}{||x||} \leq ||A|| \cdot ||A^{-1}|| \cdot \frac{||\delta f||}{||f||}$$

Величину $||A|| \cdot ||A^{-1}|| = condA$ называют числом обусловленности матрицы A. $condA \geq 1$ в любой норме, поскольку

$$E = AA^{-1} \Rightarrow ||E|| \leq ||A|| \cdot ||A^{-1}||$$

Мы получили зависимость относительной погрешности решения от возмущения входных данных через число обусловленности.

Рассмотрим случай когда $\delta f = 0$, тогда

$$\delta x = -A^{-1}\delta A \cdot x$$

$$||\delta x|| \leq ||A^{-1}|| \cdot ||\delta A|| \cdot ||x||$$
$$\frac{||\delta x||}{||x||} \leq ||A|| \cdot ||A^{-1}|| \cdot \frac{||\delta A||}{||A||}$$
$$\frac{||\delta x||}{||x||} \leq condA \cdot \frac{||\delta A||}{||A||}$$

Запишем в качестве окончательного результата обобщенный третий случай для малых возмущений $||A^{-1}|| \cdot ||\delta A|| \ll 1$

$$\frac{||\delta x||}{||x||} \leq \frac{condA}{1 - ||A^{-1}|| \cdot ||\delta A||}\left(\frac{||\delta f||}{||f||} + \frac{||\delta A||}{||A||}\right) =$$
$$= \frac{condA}{1 - condA\frac{||\delta A||}{||A||}}\left(\frac{||\delta f||}{||f||} + \frac{||\delta A||}{||A||}\right)$$

Проведем анализ полученной формулы, посмотрим какую роль вносят погрешности округления $\frac{||\delta f||}{||f||}$

$$\frac{||\delta A||}{||A||} \sim \frac{||\delta f||}{||f||} \sim O(n\varepsilon_M)$$

Погрешность решения связана с погрешностями округления как

$$\frac{||\delta x||}{||x||} = O(condA \cdot n \cdot \varepsilon_M)$$

Вспомним понятие нормы в конечномерном случае

1) Евклидова норма

$$||x|| = \sqrt{(x, x)} = \sqrt{|x_1|^2 + \ldots + |x_n|^2}$$

модуль нужен для того, чтобы формула работала и в комплексном случае.

2)p-норма

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

3) Равномерная или c-норма

$$||x||_c = \max_i |x_i| = ||x||_\infty = \lim_{p\to\infty} ||x||_p$$

Матричные нормы согласованы с нормой вектора

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

$$\frac{||Ax||}{||x||} \leq ||A||$$

$$||Ax|| \leq ||A|| \cdot ||x||$$

Тогда c-норма

$$||A||_c = \max_i \left(\sum_{j=1}^{n} |a_{ij}| \right)$$

Евклидова или сферическая норма

$$||A||_E = \left(\sum_{i,j} |a_{ij}|^2 \right)^{\frac{1}{2}}$$

Спектральная норма

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

Максимальная норма

$$||A||_m = n \max_{i,j} |a_{ig}|$$

Задание. Показать согласованность $||A||_s$ и $||A||_2$ и согласованность $||A||_m$ с $||A||_c, ||A||_1, ||A||_2$.

В R^n имеет место соотношение

$$||x||_1 \leq ||x||_2 \leq ||x||_c \leq \sqrt{n}||x||_2 \leq n||x||_1$$

т.е в R^n все эти нормы являются согласованными, сходимость в одной из них влечет сходимость во всех остальных.

Для доказательства первого неравенства воспользуемся неравенством Коши-Буняковского

$$\sum_{i=1}^{n} |x_i| \cdot 1 \leq \left(\sum_{i=1}^{n} |x_i|^2 \right)^{\frac{1}{2}} \left(\sum_{i=1}^{n} 1^2 \right)^{\frac{1}{2}} = \frac{(\sqrt{n})^2}{\sqrt{n}} \left(\sum_{i=1}^{n} |x_i|^2 \right)^{\frac{1}{2}} \Rightarrow$$

$$\frac{1}{n} \sum_{i=1}^{n} |x_i| \leq \left(\frac{1}{n} \sum_{i=1}^{n} |x_i|^2 \right)^{\frac{1}{2}}$$

Задание. Предлагается поразмышлять над остальными неравенствами самостоятельно.

8.1. Метод последовательного исключения Гаусса

Запишем снова наше основное уравнение

$$Ax = f$$

Напишем в развернутом виде систему

$$a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = f_1$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$a_{k1}x_1 + a_{k2}x_2 + \ldots + a_{kn}x_n = f_k$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = f_n$$

Метод последовательного исключения Гаусса состоит в том, чтобы по очереди исключать переменные x_k до тех пор, пока не получится свести исходную задачу к равносильной задаче

$$Ax = f \Leftrightarrow Ux = y$$

Где U верхнетреугольная матрица с единицами на диагонали

$$(U)_{ij} = \begin{cases} 1, & j = 0 \\ u_{ij}, & j > i \\ 0, & j \end{cases}$$

Получение этой системы, т.е. построение матрицы U и вектора y составляют, так называемый, прямой ход метода исключения Гаусса. Дальнейшее решение системы $Ux = y$ является обратным ходом метода исключения.

Прямой ход.

Пусть $a_{11} \neq 0$. Тогда, деля первое уравнение на a_{11}, получим

$$x_1 + u_{12}x_2 + \ldots + u_{1n}x_n = y_1$$

$$u_{1k} = \frac{a_{1k}}{a_{11}}; \quad k = \overline{2, n}; \quad y_1 = \frac{f_1}{a_{11}}$$

Комбинируя полученное уравнение с оставшимися уравнениями системы, исключим в них переменную x_1, получим

$$0 \cdot x_1 + a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + \ldots + a_{2n}^{(1)}x_n = f_2^{(1)}$$

$$0 \cdot x_1 + a_{32}^{(1)}x_2 + a_{33}^{(1)}x_3 + \ldots + a_{3n}^{(1)}x_n = f_3^{(1)}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$0 \cdot x_1 + a_{n2}^{(1)}x_2 + a_{n3}^{(1)}x_3 + \ldots + a_{nn}^{(1)}x_n = f_n^{(1)}$$

Для оставшихся уравнений (без x_1) повторим описанную процедуру, тогда на шаге S (после $s-1$ шага исключения) имеем

$$a_{s,s}^{(s-1)}x_s + a_{s,s+1}^{(s-1)}x_{s+1} + \ldots + a_{s,n}^{(s-1)}x_n = f_s^{(s-1)}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$a_{i,s}^{(s-1)}x_s + a_{i,s+1}^{(s-1)}x_{s+1} + \ldots + a_{i,n}^{(s-1)}x_n = f_i^{(s-1)} \tag{36}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$a_{n,s}^{(s-1)}x_s + a_{n,s+1}^{(s-1)}x_{s+1} + \ldots + a_{n,n}^{(s-1)}x_n = f_n^{(s-1)}$$

Это редуцированная система относительно пока не исключенных переменных $x_s, x_{s+1}, \ldots, x_n$.

Положив $a_{s,s}^{(s-1)} \neq 0$, поделим на него s-ое уравнение

$$x_s + u_{s,s+1}x_{s+1} + \ldots + u_{s,n}x_n = y_s$$

$$u_{s,j} = \frac{a_{s,j}^{(s-1)}}{a_{s,s}^{(s-1)}}; \quad j = s+1, \ldots, n; \quad y_s = \frac{f_s^{(s-1)}}{a_{s,s}^{(s-1)}}$$

Умножим полученное уравнение $a_{i,s}^{(s-1)}$ и вычтем из i-уравнения, исключив величину x_s в системе (36), получим

$$x_s + u_{s,s+1}x_{s+1} + \ldots + u_{s,n}x_n = y_s$$

$$a_{s+1,s+1}^{(s)}s_{s+1} + \ldots + a_{s+1,u}^{(s)}x_n = f_{s+1}^{(s)}$$

$$a_{n,s+1}^{(s)}x_{s+1} + \ldots + a_{n,n}^{(s)}x_n = f_n^{(s)}$$

где

$$a_{ij}^{(s)} = a_{i,j}^{(s-1)} - a_{i,s}^{(s-1)}u_{s,j}; \quad i,j = \overline{s+1,n}$$

$$f_i^{(s)} = f_i^{(s-1)} - a_{i,s}^{(s-1)}y_s; \quad i,j = \overline{s+1,n}$$

Таким образом, прямой ход в методе Гаусса

$$Ax = F \Leftrightarrow Ux = y$$

вычисляется по формулам

$$u_{s,j} = \frac{a_{s,j}^{(s-1)}}{a_{s,s}^{(s-1)}}, \quad s = 1, \ldots, n; \;\; j = s+1, \ldots, n$$

$a_{i,j}^{(s)} = a_{i,j}^{(s-1)} - a_{i,s}^{(s-1)}u_{s,j}; \quad i,j = s+1, \ldots, n; \;\; s = 1, \ldots, n-1$

Для матрицы и для правой части по формулам:

$$y_s = \frac{f_s^{(s-1)}}{a_{s,s}^{(s-1)}}; \quad s = 1, \ldots, n; \;\; f_s^{(0)} = f_s$$

$$f_i^{(s)} = f_i^{(s-1)} - a_{i,s}^{(s-1)} y_s; \;\; i = s+1, \ldots, n; \;\; s = 1, \ldots, n-1$$

Обратный ход метода Гаусса. Теперь решим систему $Ux = y$ с верхнетреугольной матрицей с единицами в качестве диагональных элементов

$$\begin{cases} x_n = y_n \\ x_i = y_i - \sum_{j=i+1}^{n} u_j x_j, \quad i = \overline{n-1, 1} \end{cases}$$

8.2. LU - разложение невырожденной матрицы

Запишем нашу задачу $Ax = f \Leftrightarrow Ux = y$ в другом виде

$$Ly = f; \quad Ux = y$$

$$A = LU; \quad LUx = f$$

С помощью матрицы L строим промежуточное представление y, а затем с помощью матрицы U решаем уравнение $Ux = y$. Для того чтобы решить задачу в таком виде докажем теорему о возможности представления матриц A в виде произведения двух треугольных матриц LU

Теорема 10. LU- разложение

Пусть все главные миноры матрицы A не равны нулю, $\triangle_i(A) \neq 0$ при $i = \overline{1, n}$, тогда существует единственное представление матрицы A в виде

$$A = LU$$

где L нижнетреугольная матрица

$$(L)_{ij} = \begin{cases} l_{ij}, & j \leq i \\ 0, & j > i \end{cases}$$

U верхнетреугольная матрица с диагональными элементами равными единице

$$(U)_{ij} = \begin{cases} 1, & j = 0 \\ u_{ij}, & j > i \\ 0, & j \end{cases}$$

Доказательство.

Будем проводить доказательство по индукции. При $n = 1$ разложение очевидно

$$a_{11} = (a_{11}) \cdot (1)$$

Будем считать, что для матрицы размера $(s-1 \times s-1)$ разложение существует. Докажем что теорема верна для $n = s$

$$\left(\begin{array}{ccccc|c} & & & & & a_{1,s} \\ & & & & & a_{2,s} \\ & A_{s-1} & & & & \vdots \\ & & & & & a_{s-1,s} \\ \hline a_{s,1} & a_{s,2} & \ldots & a_{s,s-1} & & a_{s,s} \end{array}\right)$$

Введем обозначения

$$B_{1\times(s-1)} = (a_{s,1}, \dots, a_{s,s-1})$$
$$C_{(s-1)\times 1} = (a_{1,s}, \dots, a_{s-1,s})^T$$

Нам нужно доказать справедливость разложения

$$A_s = \left(\begin{array}{ccc|c} & & & 0 \\ & L_{s-1} & & 0 \\ & & & \vdots \\ & & & 0 \\ \hline l_{s,1} & l_{s,2} & \dots\ l_{s,s-1} & l_{s,s} \end{array}\right) \left(\begin{array}{ccc|c} & & & u_{1,s} \\ & U_{s-1} & & u_{2,s} \\ & & & \vdots \\ & & & u_{s-1,s} \\ \hline 0 & 0 & \dots\ \dots\ \dots\ 0 & 1 \end{array}\right)$$

Введем обозначения

$$l_{1\times(s-1)} = (l_{s,1}, \dots, l_{s,s-1})$$
$$l_{(s-1)\times 1} = (u_{1,s}, \dots, u_{s-1,s})^T$$

Запишем конечную систему и ее фрормальное решение

$$\begin{cases} L_{s-1}U_{s-1} = A_{s-1} & \\ L_{s-1}U = C; & U = L_{s-1}^{-1}C \\ l \cdot U_{s-1} = B; & l = BU_{s-1}^{-1} \\ lu + l_{ss} \cdot 1 = a_{ss} & \end{cases} \Rightarrow A = LU \tag{37}$$

Запись решения в системе (37) символическая, речь идет не о поиске обратной матрицы, а о том, чтобы решить систему уравнений с нижнетреугольной и верхнетреугольной матрицей. L_s - невырожденная матрица, покажем это

$$detA_s = \triangle_s(A) \neq 0 = detL_s \cdot detU_s = l_{s,s} \underbrace{detU_{s-1}}_{\equiv 1} detL_{s-1}$$

Тем самым получаем, что

$$l_{s,s} \neq 0$$

Мы доказали существование LU-разложения, теперь докажем единственность. Будем действовать от противного, считая что существует еще одно разложение

$$\tilde{L}\tilde{U} = \overline{L}\,\overline{U}$$
$$\overline{L}^{-1}\tilde{L} = \overline{U}\tilde{U}^{-1}$$

Здесь $\overline{L}^{-1}\tilde{L}$ нижнетреугольная матрица, а $\overline{U}\tilde{U}^{-1}$ верхнетреугольная матрица (с единичной диагональю). Так как эти матрицы равны, то отсюда мы делаем вывод, что эти матрицы являются диагональными. А так как элементы диагонали матрицы $\overline{U}\tilde{U}^{-1}$ равны единице, то матрицы являются единичными.

$$\overline{L}^{-1}\tilde{L} = \overline{U}\tilde{U}^{-1} = E \Rightarrow$$
$$\overline{L} = \tilde{L}, \quad \overline{U} = \tilde{U}$$

Следовательно разложение единственно.

Замечание. В методе Гаусса мы предполагали, что все главные миноры $\triangle_i$ матрицы A отличны от нуля. Сформулируем теорему, которая поможет нам обосновать наше предположение.

Теорема 11.

Для любой матрицы A, определитель которой не равен нулю, существует (но не единственна) матрица перестановок P такая, что при перестановке строк матрицы A главные миноры отличны от нуля

$$\triangle_i(PA) \neq 0$$

То есть для любой невырожденной матрицы можно указать такую перестановку строк, что у новой матрицы главные миноры будут отличны от нуля, тем самым мы можем применить метод Гаусса. Аналогичную по смыслу теоремы можно сформулировать и для столбцов матрицы.

Вернемся к решению нашей задачи $Ax = y$, которую опираясь на выше доказанные теоремы, мы смогли свести к задаче

$$L(Ux) = f$$

Матрицы L и U считаются известными
1) На первом этапе решаем уравнение

$$Ly = f$$

с невырожденной нижнетреугольной матрицей L и строим вектор y. Выпишем формулы для k-го уравнения

$$l_{k,1}y_1 + l_{k,2}y_2 + \ldots + l_{k,k-1}y_{k-1} + l_{k,k}y_k = f_k$$

$$y_k = \frac{f_k - \sum_{j=1}^{k-1} l_{k,j}y_j}{l_{k,k}}, \qquad k = \overline{2,n}$$

2) Далее решаем систему с верхнетреугольной матрицей, начиная с последнего уравнения системы, так как оно содержит только одну переменную

$$Ux = y$$

$$1 \cdot x_n = y_n \qquad n\text{-ое уравнение}$$

$$1 \cdot x_k + u_{k,k+1}x_{k+1} + \ldots + u_{k,n}x_n = y_k \qquad k\text{-ое уравнение}$$

Величины, вычисленные на предыдущих этапах u_i, x_i где $i = \overline{k+1,n}$, известны, поэтому

$$x_k = y_k - \sum_{j=k+1}^{n} u_{k,j}x_j, \qquad j = \overline{k+1,n}$$

Задание. Напишите формулы для y_k для специфического вида уравнения

$$Ax = E$$

9. Лекция 9. Итерационные методы решения систем линейных уравнений. Часть 1

9.1. LU-разложение ленточной матрицы

Матрица A называется ленточной с шириной ленты k, если

$$(A)_{i,j} \begin{cases} a_{ij}, & |i-j| \le k \\ 0, & |i-j| > k \end{cases}$$

Ширина ленты $(2k+1)$ элемент. Если взять $k=1$, тогда мы получим трехдиагональную матрицу. Удобства работ с ленточными матрицами объясняется прежде всего компактностью способа их хранения. Требуется хранить не более $n \cdot (2k+1)$ элементов (даже меньше), а не n^2 как в обычном случае. Также удобным является LU-разложение таких матриц. При работе с ленточными матрицами крайне невыгодна перестановка уравнений, поскольку при этом увеличивается ширина ленты.

Теорема 12.

Если у ленточной матрицы A существует LU-разложение, тогда матрицы L и U треугольные ленточные матрицы той же структуры.

Эту теорему мы докажем для частного случая трехдиагональной матрицы, для которой реализация LU-разложения носит название метода прогонки.

Доказательство.

Пусть у нас есть трехдиагональная матрица

$$A = \begin{pmatrix} b_1 & c_1 & 0 & 0 & \dots & 0 \\ a_2 & b_2 & c_2 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & a_{n-1} & b_{n-1} & c_{n-1} \\ 0 & 0 & 0 & \dots & a_n & b_n \end{pmatrix}$$

Запишем для такой матрицы k-ое уравнение системы $Ax = f$

$$a_k x_{k-1} + b_k x_k + c_k x_{k+1} = f_k, \qquad k = \overline{1,n}$$

$$a_n = 0$$

$$c_n = 0$$

Построим формулы LU-разложения

$$L = \begin{pmatrix} \beta_1 & 0 & 0 & 0 & \dots & 0 \\ \alpha_2 & \beta_2 & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \alpha_{n-1} & \beta_{n-1} & 0 \\ 0 & 0 & 0 & \dots & a_n & b_n \end{pmatrix}, U = \begin{pmatrix} 1 & \gamma_1 & 0 & 0 & \dots & 0 \\ 0 & 1 & \gamma_2 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 1 & \gamma_{n-1} \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}, A = LU$$

Получим k-ую строку LU разложения

$$\begin{cases} \alpha_k \cdot 1 = a_k \\ \alpha_k\gamma_{k-1} + \beta_k \cdot 1 = b_k \\ \beta_k\gamma_k = c_k \\ \beta_1 \cdot 1 = b_1 \\ \beta_1\gamma_1 = c_1 \end{cases} \Rightarrow \begin{cases} \alpha_k = a_k \\ \beta_k = b_k - \alpha_k\gamma_{k-1} \\ \gamma_k = \frac{c_k}{b_k} \\ \beta_1 \cdot 1 = b_1 \\ \beta_1\gamma_1 = c_1 \end{cases}$$

Нам нужно гарантировать корректность деления, то есть показать что $\beta_k \neq 0$. Если матрица A обладает диагональным преобладанием $|b_k| > |a_k| + |c_k| > 0$, то LU-разложение существует и единственно. Покажем, что в этом случае коэффициент $\beta_k \neq 0$

$$|\beta_k| = |b_k - \alpha_k\gamma_{k-1}| \geq |b_k| - |\alpha_k| \cdot |\gamma_{k-1}| > |a_k|(\underbrace{(1 - |\gamma_{k-1}|)}_{>0}) + |c_k|) > 0$$

Так как $|b_1| > |c_1|$ значит $|\gamma_1| < 1$, тогда по индукции

$$|\gamma_k| = \frac{|c_k|}{|\beta_k|} < \frac{|c_k|}{|a_k|(1 - |\gamma_{k-1}|) + |c_k|} < 1$$

Итак, мы построили LU-разложение. Решение системы $Ax = y$ строим в два этапа
а) *прямой ход прогонки* - находим y из $Ly = f$

$$y_1 = \frac{f_1}{\beta_1}$$

$$y_k = \frac{f_k - a_k y_{k-1}}{\beta_k}, \qquad k = \overline{2, n}$$

б) *Прямой ход прогонки* - находим x из $Ux = y$

$$x_n = y_n$$

$$x_k = y_k - \gamma_k x_{k+1}, \qquad k = \overline{n-1, 1}$$

Приведем несколько характерных оценок для нашего метода. При LU-разложении количество арифметических действий величина порядка $O(n^3)$. В частном случае для трехдиагональной и формул прогонки нам необходимо $O(n)$ действий.

Методы, которые мы рассмотрели, входят в категорию прямых методов, когда мы реализуем алгоритм, дающий результат за конечное число шагов. Вторая группа методов носит название итерационных.

9.2. Итерационные методы решения СЛАУ

Напомним еще раз, что мы решаем задачу

$$Ax = f; \qquad A : R^m \Rightarrow R^m \tag{38}$$

Одношагового итерационный метод строится по формуле

$$B_{n-1}\frac{x_{n-1}-x_n}{\tau_{n+1}}+Ax_n=f$$

где индекс n это номер итерации; B - итерационная матрица; τ - итерационный параметр.

Как мы видим, в общем случае и итерационная матрица, и итерационный параметр изменяются в зависимости от номера итерации. Мы ограничимся рассмотрением стационарных итерационных методов, где B и τ фиксированы.

$$B\frac{x_{n-1}-x_n}{\tau}+Ax_n=f\Rightarrow$$
$$B\frac{x_{n-1}-x_n}{\tau}=f-Ax_n=r_n \tag{39}$$

r_n невязка уравнения, векторная величина.

Замечание. Решая уравнение (39) на очередной итерации номер n, когда мы получим решение x_n, мы не сможем точно удовлетворить уравнению (38), то есть мы не получим точного равенства. Таким образом, мы вводим величину r_n, которая компенсирует разность точного решения уравнения и решения полученного нами на n-ой итерации.

Лучше всего в уравнении (39) выполнить очередной шаг x_{n+1}, так чтобы поправка к найденному решению x_n компенсировала невязку, возникшую на очередном шаге

$$B(x_{n+1}-x_n)=\tau r_n$$

Нам необходимо чтобы итерационная матрица была достаточно легко обратима. То есть поиск решения выше написанного уравнения должен быть значительно проще, чем поиск решения уравнения (38).

$$x_{n+1}=x_n+\tau B^{-1}r_n$$

$$r_{n+1}=f-Ax_{n+1}=\underbrace{f-Ax_n}_{r_n}-\tau AB^{-1}r_n=(E-\tau AB^{-1})r_n$$

Величина

$$e_n=x-x_n$$

называется ошибкой очередного приближения. Это разность точного решения и решения на n-ом шаге

$$r_n=f-Ax_n=Ax-Ax_n=Ae_n$$

Тем самым мы получили связь между невязкой и погрешностью решения.

$$Ae_{n+1}=(E-\tau AB^{-1})Ae_n$$

$$e_{n+1}=(E-\tau A^{-1}AB^{-1}A)e_n=(E-\tau B^{-1}A)e_n$$

Мы получили итерационный процесс для погрешностей. Невязка и погрешности находятся в разных пространствах. Вернемся к формуле итерационного процесса

$$x_{n+1} = x_n + \tau B^{-1} r_n = x_n + \tau B^{-1}(f - Ax_n) = (E - \tau B^{-1}A)x_n + \tau B^{-1} f \Rightarrow$$

$$x_{n+1} = Cx_n + g \tag{40}$$

Матрицу C будем называть матрицей перехода к очередной итерации. Преобразуем уравнение (40) понизив индекс

$$x_n = Cx_{n-1} + g = C(Cx_{n-2} + g) + g = C^2 x_{n-2} + (E + C)g =$$

$$C^3 x_{n-3} + (E + C + C^2)g = \ldots = C^n x_0 + (E + C + C^2 + \ldots + C^{n-1})g$$

Таким образом, вопрос сходимости итерационного процесса сводится к изучению сходимости матричного степенного ряда. Сформулируем известные теоремы, чтобы воспользоваться их результатом.

Теорема 13

Для сходимости матричного степенного ряда

$$\sum_{k=0}^{\infty} \alpha_k C^k$$

необходимо и достаточно чтобы все собственные значения матрицы принадлежали области сходимости производящего ряда

$$\sum_{k=0}^{\infty} \alpha_k \lambda^k$$

с кругом сходимости $|\lambda| < R$, то есть

$$\forall i : |\lambda_i(C)| < R$$

Теорема 14

Для сходимости метода последовательных приближений необходимо и достаточно, чтобы все собственные значения матрицы перехода удовлетворяли бы условию

$$|\lambda_i| < 1, \quad \forall i = \overline{1, n}$$

Имея возможность оценить собственные значения λ_i через норму матрицы C, получаем достаточные условия сходимости метода последовательных приближений. Напомним, для собственного значения λ справедливо

$$Cx = \lambda x \Rightarrow \quad ||Cx|| = \lambda ||x||$$

но при этом

$$||Cx|| \leq ||C|| \cdot ||x|| \Leftrightarrow |\lambda| \leq ||C||$$

Условие $||C|| < 1$ является достаточным для сходимости метода последовательных приближений.

Необходимым условием является

$$C^n \to 0$$

Посмотрим к чему стремится частичная сумма S_n нашего матричного ряд

$$E + C + C^2 + \ldots + C^{n-1}$$

$$(E-C)^{-1}(E+C+C^2+\ldots+C^{n-1}) = E+C+C^2+\ldots+C^{n-1}-C-\ldots-C^{n-1}-C^n =$$
$$= E - C^n \to E$$

Это означает, что частичная сумма стремиться к обратной матрице $(E-C)^{-1}$
Теперь перейдем к пределу в выражении

$$C^n x_0 + (E + C + C^2 + \ldots + C^{n-1})g$$

$$\lim_{n\to\infty} x_n = (E-C)^{-1}g = (\tau B^{-1}A)^{-1}\tau B^{-1}f = A^{-1}B\frac{1}{\tau}\tau B^{-1}f = A^{-1}f = x$$

Итерационная последовательность, при выполнении условия $||C|| < 1$ для любого x_0, сходится к решению задачи

$$\begin{cases} x_{n+1} = Cx_n + g \\ x_0 = \overset{(0)}{x} \end{cases}$$

9.3. Основные итерационные методы

Метод релаксации
Итерационная матрица $B = E$; параметр $\tau > 0$

$$\frac{x_{n+1} - x_n}{\tau} + Ax_n = f$$

или учитывая что $r_n = f - Ax_n$

$$\frac{x_{n+1} - x_n}{\tau} = r_n$$

$$(x_{n+1})_i = (x_n)_i + \tau(r_n)_i; \quad i = \overline{1,n}$$

Следующие методы связаны со специфическим разбиением матрицы A, которое можно записать в следующем виде

$$A = (A_L + D + A_U)$$

$D = diag(a_{11}, \ldots, a_{mm})$ - диагональная матрица, состоящая из элементов главной диагонали a_{ii} матрицы A.

A_L - нижнетреугольная матрица, состоящая из элементов матрицы A стоящих ниже главной диагонали.

$$(A_L)_{ij} = \begin{cases} a_{ij}, & j0, \\ & j \geq i \end{cases}$$

A_L - верхнетреугольная матрица, состоящая из элементов матрицы A стоящих выше главной диагонали.

$$(A_U)_{ij} = \begin{cases} a_{ij}, & j > i \\ 0, & j \leq i \end{cases}$$

Метод Якоби
Итерационная матрица $B = D$, итерационный параметр $\tau = 1$

$$D(x_{n+1} - x_n) = r_n$$

$$x_{n+1} = x_n + D^{-1} r_n$$

$$(x_{n+1})_i = (x_n)_i + \frac{1}{a_{ii}}(r_n)_i$$

Метод Зейделя
Итерационная матрица $B = D + A_L$, итерационный параметр $\tau = 1$ Уравнение, которое необходимо решить

$$(D + A_L)(x_n + 1 - x_n) = r_n$$

$$x_{n+1} = x_n + (D + A_L) r_n$$

Метод SOR(Successive over-relaxation).Метод верхней релаксации
Итерационная матрица $B = D + \omega A_L$, итерационный параметр ω

$$(D + \omega A_L)\frac{x_{n-1} - x_n}{\omega} = f - Ax_n$$

$$x_{n-1} = x_n + (D + \omega A_L)^{-1} r_n$$

10. Лекция 10. Итерационные методы решения систем линейных уравнений. часть 2

Пусть в уравнении $Ax = f$ матрица A является SPD-матрицей. Напомним, что для положительно определенной матрицы справедливо

$$A > 0 : \quad (Ax, x); \quad \forall x \neq 0$$

Из этого следует (в конечномерном случае)

$$(Ax, x) \geq \delta(x, x)$$

$$A \geq \delta E$$

E-квадрат евклидовой нормы. Посмотрим, откуда берется это неравенство. Так как для SPD-матрицы $A^T = A$, тогда $\delta = \lambda_{min}$, и ее собственные значения строго больше нуля

$$(Ax, x) = \left(A \sum_i \alpha_i x_i, \sum_j \alpha_j x_j \right) = \left(\sum_i \alpha_i \lambda_i x_i, \sum_j \alpha_j x_j \right) =$$

$$= \sum_i \alpha_i^2 \lambda_i \geq \lambda_{min} \sum \alpha_i^2 = \lambda_{min}(x, x)$$

Если же матрица не симметрична, но все так же положительна, тогда мы можем сравнительно легко ее симметризовать

$$(Ax, x) = \frac{1}{2}[(Ax, x) + (x, A^* x)] = \left(\frac{A + A^*}{2} x, x \right)$$

Тем самым мы рассматриваем новый оператор $B = \frac{A+A^*}{2}$ и уже у симметричной формы выбираем

$$\delta = \lambda_{min} \frac{A + A^*}{2}$$

Итак, мы видим, что если матрица A положительна, значит что она положительно определенная. Всего этого достаточно, чтобы считать матрицу A обратимой, ведь если предположить, что вектор x лежит в ядре матрицы, то

$$0 = (Ax, x) = \delta(x, x) \neq 0 \quad \exists A^{-1}$$

тем самым приходим к противоречию.

Введем скалярное произведение

$$(x, y)_A = (Ax, y) = (x, Ay) = (Ay, x) = (y, x)_A$$

Оно порождает A-энергетическую норму

$$||x||_A = \sqrt{(Ax, x)}$$

10.1. Теорема Самарского

Введение A-энергетической нормы позволяет нам по-другому взглянуть на сходимость итерационных методов.

Теорема. (Самарского)

Пусть матрица A является SPD-матрицей, параметр $\tau > 0$, матрица $B > 0$ и $B - \frac{\tau A}{2} > 0$, тогда итерационная последовательность

$$x_{n+1} = (E - \tau B^{-1}A)x_n + \tau B^{-1}f$$

сходится в среднеквадартичной метрике к решению задачи $Ax = f$

Доказательство

Так как

$$x = x_n + e_n$$

Значит нам необходимо доказать

$$||e_n|| = \sqrt{(e_n, e_n)} \to 0$$

Оценим e_{n+1}

$$e_{n+1} = (E - \tau B^{-1}A)e_n$$

Для этого рассмотрим его A-энергетическую норму

$$||e_{n+1}||_A^2 = (Ae_{n+1}, e_{n+1}) = (Ae_n - \tau AB^{-1}Ae_n, e_n - \tau B^{-1}Ae_n) =$$
$$= (Ae_n, e_n) - \tau(Ae_n, B^{-1}Ae_n) - \tau(e_n, AB^{-1}Ae_n) + \tau^2(AB^{-1}Ae_n, \underbrace{B^{-1}Ae_n}_{w_n}) =$$

$$= ||e_n||_A^2 - 2\tau(Bw_n, w_n) + \tau^2(Aw_n, w_n)$$

$$||e_{n+1}||_A^2 = ||e_n||_A^2 - 2\tau\left(\left(B - \frac{\tau A}{2}\right)w_n, w_n\right)$$

Тем самым мы видим, что последовательность A-энергетических норм не возрастает и ограничена нулем снизу. Из этого получаем, что последовательность $||e_n||$ сходится. Далее, воспользовавшись тем, что

$$\left(\left(B - \frac{\tau A}{2}\right)w_n, w_n\right) \geq 2\tau\delta(w_n, w_n)$$

Запишем

$$-(||e_{n+1}||_A^2 - ||e_n||_A^2) + 2\tau\delta(w_n, w_n) \geq 0$$

В силу сходимости последовательности A-энергетических норм, по теореме о двух милиционерах, сходится и последовательность

$$||w_n||^2 = (w_n, w_n) \to 0 \quad n \to \infty$$

так как $w_n = B^{-1}Ae_n$, тогда

$$e_n = A^{-1}Bw_n$$

$$||e_n|| = \sqrt{(e_n, e_n)} = ||A^{-1}Bw_n|| \leq ||A^{-1}|| \cdot ||Bw_n|| \Rightarrow$$
$$||e_n|| \to 0 \quad n \to \infty$$

Теорема доказана.

10.2. Достаточные условия сходимости простейших итерационных методов

Метод Релаксации(или Ричардсона)

Матрица $B = E$

$$B - \frac{\tau A}{2} > 0 \tag{41}$$

Для SPD-матрицы

$$||A|| = \sup_{x \neq 0} = \frac{(Ax, x)}{(x, x)} \Rightarrow (Ax, x) \leq ||A||(x, x)$$

Отсюда, как мы видим, выполнено неравенство

$$A \leq ||A|| \cdot E$$

А значит неравенство (41) выполнено всегда при

$$\frac{A}{||A||} - \frac{\tau A}{2} > 0$$

Отсюда следует что

$$0 < \tau \leq \frac{2}{||A||}$$

Метод верхней релаксации(SOR)

Матрица $B = D + \omega A_L$, параметр $\tau = \omega$

$$B - \frac{\tau A}{2} = D + \omega A_L - \frac{\omega}{2}(A_L + D + A_u) = \left(1 - \frac{\omega}{2}\right)D + \frac{\omega}{2}(A_L - A_u) > 0$$

$$\left(\left(B - \frac{\tau A}{2}\right)x, x\right) = \left(1 - \frac{\omega}{2}\right)(Dx, x) + \frac{\omega}{2}(A_L x, x) - \frac{\omega}{2}(A_u x, x) = \left(1 - \frac{\omega}{2}\right)(Dx, x) > 0$$

В силу симметрии матрицы A: $((A_L x, x) - (A_u x, x)) \equiv 0$, так как $(A_L x, x) = x, (A_u^* x)$. Далее, из условия $A > 0$ следует, что матрица $D = diag(a_{11}, \ldots, a_{nn})$. Действительно, если взять в качестве вектора x базисный вектор e, тогда

$$\vec{x} = \vec{e}_i = (\underbrace{0, \ldots, 1, \ddots, 0}_{\text{1 на i-ом месте}}); \;\; (Ax, x) = a_{ii} = (Dx, x) > 0; \;\; \forall x \neq 0$$

Это означает, что все диагональные элементы матрицы больше нуля и, следовательно, матрица D положительная. Окончательно получим

$$0 < \omega < 2$$

В частности это подходит для $\omega = 1$, что было связано с методом Зейделя.

Метод Якоби

Матрица $B = D$, параметр $\tau = 1$

$$B - \frac{\tau A}{2} > 0 \Leftrightarrow D - \frac{A}{2} > 0$$

Тем самым мы должны проверить условие

$$A < 2D$$

Сформулируем достаточные условия сходимости метода Якоби.

Теорема 4. Если симметричная положительно определенная матрица с диагональным преобладанием

$$a_{ii} > \sum_{i \neq j} |a_{ii}|$$

то метод Якоби сходится (в среднеквадратичной метрике).

Доказательство

$$(Ax, x) = \sum_{i,j} a_{ij} x_i x_j \leq \sum_{i,j} |a_{ij}| \cdot |x_i| \cdot |x_j| \leq \sum_{i,j} |a_{ij}| \frac{|x_i|^2 + |x_j|^2}{2} =$$

$$= \frac{1}{2} \left(\sum_{i,j} |a_{ij}| \cdot |x_i|^2 + \sum_{i,j} |a_{ij}| \cdot |x_j|^2 \right) = \sum_{i,j} |a_{ij}| \cdot |x_i|^2 =$$

$$= \sum_i |x_i|^2 \left(a_{ii} + \sum_{i \neq j} |a_{ij}| \right) < \sum_i x_i^2 2a_{ii} = 2(Dx, x)$$

Окончательно получаем, что

$$(Ax, x) = 2(Dx, x)$$

следовательно, метод Якоби сходится

Рассмотрим сходимость метода релаксаций. Итерационный процесс имеет вид

$$x_{n+1} - x_n = \tau r_n$$

Для погрешностей

$$e_{n+1} = (E - \tau A) e_n$$

$$||e_{n+1}|| \leq ||E - \tau A|| \cdot ||e_n||$$

Покажем, что при определенных условиях можно получить оценку

$$||E - \tau A|| \leq q < 1$$

Тогда мы сможем сказать, что этот процесс, по крайней мере, первого порядка сходимости. Предположим, что нам известны границы спектра матрицы A

$$Sp(A) \subset [m, M], \quad \lambda(A) \in [m, M]$$

$$|| \underbrace{E - \tau A}_{M} || = \left| \sqrt{\max_{\lambda(A)} \lambda(MM^T)} \right| = \max_{\lambda(A)} \left| \lambda(E - \tau A) \right| =$$

Проведем промежуточные выкладки

$$Ax = \lambda x$$

$$-\tau Ax = -\tau\lambda x$$

$$(E - \tau A)x = (1 - \tau A)x$$

тогда

$$= \max_{\lambda(A)} |1 - \tau A|$$

Получив такую оценку для нормы, мы можем выбирать итерационный параметр оптимальным образом

$$\tau_0 : \quad \underbrace{argmin}_{\tau>0} \left(\max_{\lambda\in[m,M]} |1 - \tau\lambda| \right)$$

11. Лекция 11. Алгебраическая проблема поиска собственных значений

Мы будем решать задачу

$$Ax = \lambda x, \quad x \neq 0 \tag{42}$$

Мы хорошо знаем прямые методы решения этой задачи, на первом этапе ищем корни характеристического уравнения

$$P(\lambda) \equiv P_n(\lambda) = det(a - \lambda E) = 0 \tag{43}$$

Далее мы строим собственные вектора

$$A - \lambda_k E x_k = 0$$

Наряду с матрицей A рассматривают сопряженную матрицу $\overline{A^T} = A^*$. Для этой матрицы также решается задача

$$A^* y = \tilde{\lambda} y$$

Мы также знаем что $\overline{\lambda} = \tilde{\lambda}$. Если мы рассматриваем не комплексно сопряженные значения λ, тогда

$$y \perp x, \quad \overline{\lambda_k} = \tilde{\lambda_k}$$

11.1. Устойчивость невырожденной задачи нахождения собственных векторов и собственных значений

Мы ограничимся рассмотрением случая, когда все λ_k имеют кратность 1, то есть тот случай, когда с помощью собственных векторов можно построить базис. Возмущенная погрешностями задача (42) имеет вид:

$$\begin{aligned}(A + \delta A)(x_k + \delta x_k) &= (\lambda_k + \delta\lambda_k)(x_k + \delta x_k) \\ A\delta x_k + \delta A x_k &= \lambda_k \delta x_k + \delta\lambda_k x_k \end{aligned} \tag{44}$$

Слагаемые в нулевом порядке по соответствующим возмущения удовлетворяют задаче и поэтому аннулировались. Покажем, что малое возбуждение входных данных приводит к малым погрешностям решения. Будем считать, что все векторы нормированы и

$$||x_k||^2 = (x_k, x_k) = 1$$

Варьируя это равенство получим

$$(\delta x_k, x_k) = 0$$

Тогда разложение вектора δx_k по невозмущенному базису из собственных векторов $\{x_i\}$ коэффициент $a_{kk} = 0$

$$\delta x_k = \sum_{i \neq k} \alpha_{ik} x_i = \sum_i{}' \alpha_{ik} x_i$$

Штрих у суммы означает что сумма с дефектом по i (индекс i пропущен).

Теперь умножим (44) скалярно (т.е. справа) на собственный вектор y_l сопряженной матрицы A^*, получим:

1) При $l = k$, y_k - собственный вектор для $\overline{\lambda_k}$

$$\underbrace{\left(\sum_{i=1}^{n}{}' \alpha_{ki}\vec{x}_i, y_k\right)}_{\sum_{i\neq k}{}'\alpha_{ki}\vec{x}_i, y_k=0} + (\delta A x_k, y_k) = \underbrace{\lambda_k \sum_{i=1}^{n}{}' \alpha_{ki}\vec{x}_i, y_k}_{\equiv 0} + \delta\lambda_k(x_k, y_k)$$

$$(\delta A x_k, y_k) = \delta\lambda_k(x_k, y_k)$$

или

$$|\delta\lambda_k| \leq \frac{|(\delta A x_k, y_k)|}{|(x_k, y_k)|} \leq \frac{||\delta A x_k|| \cdot ||y_k||}{|(x_k, y_k)|} \leq \frac{||\delta A|| \cdot ||x_k|| \cdot ||y_k||}{|(x_k, y_k)|} \leq$$

$$\leq \max_{i,j}|\delta a_{i,j}| \frac{\sqrt{(x_k, x_k)(y_k, y_k)}}{|(x_k, y_k)|} \equiv \chi_{k,k} \max|\delta a_{i,j}|$$

Здесь χ - k-ый главный коэффициент перекоса матрицы A.

2) Аналогично при $l \neq k$

$$\underbrace{\left(\sum_{i=1}^{n}{}' \alpha_{ki}\vec{x}_i, y_l\right)}_{\text{остается только } i=l} + (\delta A x_k, y_l) = \lambda_k \sum_{i=1}^{n}{}' \alpha_{ki}\vec{x}_i, y_l + \underbrace{\delta\lambda_k(x_k, y_l)}_{\equiv 0, x_k \perp y_k}$$

$$\alpha_{kl}\lambda_l(x_l, y_l) + (\delta A x_k, y_l) = \lambda_k \alpha_{kl}(x_l, y_l)$$

отсюда получаем

$$\alpha_{lk}(\lambda_k - \lambda_l)(x_l, y_l) = (\delta A x_k, y_l)$$

теперь мы можем получить оценку коэффициентов a_{kl}

$$|a_{kl}| \leq \frac{|(\delta A x_k, y_l)|}{|(\lambda_k - \lambda_l)| \cdot |(x_l, y_l)|} \leq |\frac{\max_{ij}|\delta a_{ij}|}{|(\lambda_k - \lambda_l)|} \frac{\sqrt{(x_k, x_k)(y_l, y_l)}}{|(x_l, y_l)|} =$$

$$\frac{\chi}{|(\lambda_k - \lambda_l)|} \max \alpha_{ij}|$$

Итак:

1) собственное значение λ_k матрицы A устойчиво относительно возмущений матрицы, если соответствующий ему коэффициент перекоса χ_k мал;

2) Для устойчивости собственных векторов относительно возмущений матрицы A необходимо, чтобы все $\chi_{k,l}$ были малы.

11.2. Метод парабол

Далее нам необходимо решить характеристическое уравнение (43). Для этого воспользуемся методом парабол поскольку он может обеспечить сходимость к комплексному корню характеристического уравнения (43). Построим интерполяционный многочлен второго порядка $N_2(\lambda)$. Из уравнения $N_2(\lambda) = 0$ построим корень $\overset{(n+1)}{\lambda}$. Многочлен $N_2(\lambda) = 0$ строится, считая что нам известны три последних шага итерации, то есть известны значения $\overset{(n)}{\lambda}, \overset{(n-1)}{\lambda}, \overset{(n-2)}{\lambda}$, тогда

$$N_2(\lambda) = P(\overset{(n)}{\lambda}) + \underbrace{(\lambda - \overset{(n)}{\lambda})}_{z} P(\overset{(n)}{\lambda}, \overset{(n-1)}{\lambda}) + (\lambda - \overset{(n)}{\lambda})(\lambda - \overset{(n-1)}{\lambda}) P(\overset{(n)}{\lambda}, \overset{(n-1)}{\lambda}, \overset{(n-2)}{\lambda})$$

Получаем квадратное уравнение относительно z

$$Az^2 + Bz + C = 0$$

Постарайтесь в качестве самостоятельного упражнения решить полученное уравнение. После нахождения корней выбираем минимальный из них z_{min}. И окончательно корень $\overset{(n+1)}{\lambda}$ строим по формуле

$$N_2(\lambda) = 0 \rightarrow \quad \overset{(n+1)}{\lambda} \pm \overset{(n)}{\lambda} + z_{min} \tag{45}$$

Для того чтобы решить задачу таким образом, нам необходимо $O(n^4)$ действий. Это технически сложно, поэтому на практике стараются заменить эту процедуру более удобной. Сформулируем теорему

Теорема.(Шура)

Для любой матрицы A существует невырожденная матрица P такая, что матрица $B = P^{-1}AP$ является трехдиагональной матрицей

$$(B)_{ij} = \begin{cases} b_i; & |i-j| \leq 1 \\ 0; & |i-j| > 0 \end{cases}$$

Тогда, если собственные значения и собственные векторы матрицы B известны, то

$$By = \alpha y \Leftrightarrow P^{-1}AP = \alpha y \Leftrightarrow A(Py) = \alpha(Py)$$

или

$$Ax = \alpha x$$

Таким образом, собственные значения исходной матрицы A и матрицы B совпадают $\lambda_i = \alpha_i$. А собственный вектор x_i, отвечающий данному значению λ_i, строится через собственный вектор y_i

$$x_i = Py_i$$

Тем самым мы можем перейти к задаче нахождения собственных векторов и собственных значений матрицы B. Начнем с нахождения собственных значений

матрицы B, т.е. корней характеристического многочлена $P_B(a) = 0$. Для метода интерполяции необходимо вычисление определителя $det(B - \alpha E) = D_n$ в узлах сетки $\{\alpha_i\}$. Эти вычисления можно реализовать по экономичной расчетной схеме, что весьма важно.

$$A = \begin{pmatrix} & \ddots & & \ddots & 0 & 0 \\ \ddots & & \ddots & & & \\ \vdots & \ddots & & \ddots & \ddots & 0 \\ \vdots & \vdots & \ddots & & \ddots & \ddots \\ 0 & 0 & \dots & b_{m-1,m-2} & b_{m-1,m-1-\alpha} & b_{m-1,m} \\ 0 & 0 & 0 & \dots & b_{m,m-1} & b_{n,m-\alpha} \end{pmatrix} =$$

$$= (b_{m,m-\alpha})D_{m-1}(\alpha) - b_{m,m-1}bm-1,mD_{m-2}(\alpha)$$

Итак, мы получили рекуррентные формулы вычисления характеристического многочлена 3-х диагональной матицы

$$D_m(\alpha) = (b_{m,m-\alpha})D_{m-1}(\alpha) - b_{m,m-1} \cdot b_{m-1,m} \cdot D_{m-2}(\alpha)$$
$$D_0 = 1; D_{-1} = 0$$

Основная сложность метода состоит в нахождении матрицы P для приведения матрицы A к трехдиагональному виду.

Нахождение собственного вектора

Первая проблема заключается в том, что определитель матрицы $A - \lambda_k E = 0$. Вторая проблема заключается в том, что нам известно только приближенное значение $\tilde{\lambda_k}$, и тогда уравнение

$$(A - \tilde{\lambda}_k E)x = 0$$

имеет только тривиальное решения, потому что его матрица невырождена.

Решим следующую задачу

$$(A - \tilde{\lambda_k}E)x = x_0; \quad \forall x_0 \neq 0$$

Для удобства перепишем ее в виде

$$Ax - \tilde{\lambda}_k x = x_0 \tag{46}$$

Напишем разложения по базису из собственных векторов для x_0 и x

$$x_0 = \sum_i \alpha_i x_i; \quad x = \sum_i \beta_i x_i$$

Подставим полученные разложения в (46) и получим

$$\sum_i \beta_i \lambda_i x_i - \tilde{\lambda}_k \sum_i \beta_i x_i = \sum_i \alpha_i x_i$$

Приравнивая коэффициенты при одинаковых собственных векторах, окончательно имеем

$$\beta_i(\lambda_i - \tilde{\lambda}_k) = \alpha_i \;\; \Leftrightarrow \;\; \beta_i = \frac{\alpha_i}{\lambda_i - \tilde{\lambda}_k}$$

Таким образом, мы получили что коэффициент β_k достаточно большой, ведь в знаменателе дроби стоит разность реального и приближенного собственного значения. Это означает, что вектор x почти коллинеарен x_k.

Организуем итерационный процесс следующего вида

$$\begin{cases} (A - \tilde{\lambda}_k E) \overset{(n+1)}{\vec{x}_k} = \overset{(n)}{\vec{x}_k}; \\ | \overset{(n+1)}{\vec{x}} | = 1; \\ \overset{(0)}{\vec{x}} = \vec{x}_0; \\ |\overset{(0)}{\vec{x}}| = 1; \end{cases}$$

Каждый раз совершая шаг итерации вектор $\vec{x}$ нужно нормировать.

11.3. Метод вращений (Якоби)

Известно, что для действительной симметричной матрицы $A^T = A$ существует ортогональное преобразование U такое, что

$$U^T A U = diag(\lambda_1, \ldots, \lambda_n) = \Lambda$$

Для диагональной матрицы задача на собственные значения решается гораздо проще, это просто элементы стоящие на диагонали. А собственные вектора это i-ый базисный вектор.

Пускай матрица Λ построена. Тогда собственные значения найдены, так как у подобных матриц они совпадают, а собственные вектора находим как столбцы матрицы U. Итак, построим ортогональную матрицу U, которая приводила бы симметричную матрицу A к диагональному виду. В методе Якоби матрица U строится итерационно, через последовательность элементарных вращений. Рассмотрим вращение в плоскости Okl. Определим φ из условия $(\tilde{A})_{kl} = (\tilde{A})_{lk} = 0$. Найдем $(\tilde{A})_{kl}$

1) Матрица $B = AU_{kl}$ отличается от A двумя столбцами k-ым и l-ым (ведь все остальные столбцы в U единичные)

$$b_{ik} = \sum_j \alpha_{ij}(U_{kl})_{jk} = \alpha_{ik}\cos\varphi + a_{il}\sin\varphi$$

$$b_{il} = \sum_j \alpha_{ij}(U_{kl})_{jl} = \alpha_{ik}(-\sin\varphi) + a_{il}\cos\varphi$$

2) Матрица $\tilde{A} = U^T B$- отличается от B двумя строками: k-ой и l-ой

$$(\tilde{A})_{ki} = \sum_j (U^T_{kl})_{kj} b_{ji} = \cos\varphi b_{ki} + \sin\varphi b_{li}$$

$$U_{k,l}(\varphi) = \begin{pmatrix} 1 & & 0 & & & & & & & \\ & \ddots & & & & 0 & & & 0 & \\ 0 & & 1 & & & & & & & \\ & & & \cos\varphi & & & -\sin\varphi & & & \\ & & & & 1 & & 0 & & & \\ & 0 & & & & \ddots & & & 0 & \\ & & & & 0 & & 1 & & & \\ & & & \sin\varphi & & & \cos\varphi & & & \\ & & & & & & & 1 & & \\ & 0 & & & & 0 & & & \ddots & \\ & & & & & & & & & 1 \end{pmatrix}, \quad \tilde{A} = U_{kl}^T A U_{kl}.$$

Рис. 11.1

$$(\tilde{A})_{li} = \sum_j (U_{kl}^T)_{lj} b_{ji} = (-\sin\varphi) b_{ki} + \cos\varphi b_{li}$$

Окончательно

$$(\tilde{A})_{kl} = \cos\varphi b_{kl} + \sin\varphi b_{ll} = \cos\varphi(a_{kk}(-\sin\varphi) + a_{kl}\cos\varphi) + \sin\varphi(a_{lk}(-\sin\varphi) + a_{ll}\cos\varphi) =$$

$$= a_{kl}\cos 2\varphi - (a_{kk} - a_{ll})\frac{1}{2}\sin 2\varphi = 0 \;\Rightarrow\; \tan 2\varphi = \frac{2a_{kl}}{a_{kk} - a_{ll}}, \quad \varphi < \frac{\pi}{4}$$

Тем самым мы определили элементарное вращение $U_{kl}(\varphi)$.

11.4. Оценка нормы матрицы

Введем для матрицы A *сферическую норму*

$$||A||_s^2 = \sum_{ij} a_{ij}^2 = \sum_i a_{ii}^2 + \sum_{\substack{i,j \\ i \neq j}} a_{ij}^2 = S_1 + S_2$$

Покажем, что величина S не изменяется при вращении $U_{kl}(\varphi)$.

1) Преобразование $B = AU$ изменяет лишь два столбца k-ый и l-ый в матрице A, причем $b_{ik}^2 + b_{il}^2 = a_{ik}^2 + a_{il}^2$ и в S нет изменений.

2) Преобразование $\tilde{A} = U^T B$ изменяет лишь две строки в матрице B, причем $\tilde{a}_{ki}^2 + \tilde{a}_{li}^2 = b_{ki}^2 + b_{li}^2$ и в S нет изменений. Поэтому

$$||\tilde{A}||_S = ||U^T B||_S = ||B||_S = ||AU||_S = ||A||_S = inv$$

При вращении $U_{kl}(\varphi)$ вклад, который вносит сумма S_2 убывает и, следовательно, вклад S_1 возрастает. Нужно подобрать такую последовательность вращений, чтобы последовательность $S_2 \to 0$ и при этом $\tilde{A}$ станет диагональной матрицей. Для уменьшения объема вычислений поступают следующим образом:

1) Составляют суммы строк (полустрок) и находят строку с наибольшей суммой

$$S_i = \sum_{\substack{j \\ i \neq j}} a_{ij}^2 \quad \Rightarrow \quad S_{i_{max}}$$

2) В i_{max}-строке находят наибольший по модулю элемент

$$|a_{i_{max},j_{max}}|$$

3) Его исключают на очередном шаге

$$k = i_{max}, l = j_{max}$$

Тогда $S_2 \downarrow$ не менее, чем на $\frac{1}{n-1}$ от всей суммы $S_{i_{max}}$ т. е. на $\frac{1}{n-1}$ от $\frac{1}{n}S_2 \Rightarrow$ итого на долю $\frac{2}{n(n-1)}S_2$ (ибо исключаются два слагаемых). После N исключений

$$S_2^{(N)} \approx \left(1 - \frac{2}{n(n-1)}\right)^N S_2 \approx e^{-\frac{2}{n^2}N} S_2, \quad S_2^{(N)} \to 0, \quad N \to \infty$$

12. Лекция 12. Задачи минимизации

Пусть у нас есть некоторая целевая функция $\Phi(x)$ где $x \in \mathcal{X}$. Ставится задача найти минимальное значение для этой функции на множестве $\mathcal{X}$. Мы ограничимся рассмотрением тех задач, в которых минимум достигается в некоторой точке x^*

$$\Phi(x^*) = \min_{\mathcal{X}} \Phi(x)$$

тем самым x^* будет принадлежать множеству локальных минимумов функции $\Phi(x)$

$$x^* \in loc \min_{\mathcal{X}} \Phi(x)$$

Точка x^* называется *локальным минимумом* если существует такая δ-окрестность точки, что

$$\forall x: \quad \rho(x, x^*) < \delta \qquad \Phi(x) \geq \Phi(x^*)$$

Часто говорят о строгом экстремуме, тогда δ-окрестность должна быть проколотой

$$\forall x: \quad 0 < \rho(x, x^*) < \delta \quad \Phi(x) > \Phi(x^*)$$

Рассмотрим три основные задачи о минимизации функций.
1)$\mathcal{X} = R$. Задача минимизации функции одного переменного.
2)$\mathcal{X} = R^m$. Задача минимизации функции n переменных.
3))$\mathcal{X}$ - гильбертово пространство и задача о минимизации функционала.
Во всех остальных случаях мы будем рассматривать задачи без ограничений.

12.1. Минимизация функции одного переменного. Методы нулевого порядка.

Начнем с простейшего случая когда $x \in R$. Будем придерживаться обозначений $\Phi \equiv f$. Необходимо построить точку локального экстремума

$$x^* \in loc \min_{R} f(x)$$

функции $f(x)$.

Из курса матанализа нам известны теоремы о необходимых и достаточных условиях экстремума достаточно гладких функций.

Если $\Phi(x) \in C^2(R)$ тогда необходимое условие экстремума

$$\Phi'(x^*) = 0$$

производная функции в случае дифференцируемости функции равна нулю. Второе необходимое условие

$$\Phi''(x^*) \geq 0$$

Зачастую мы сталкиваемся с функциями не удовлетворяющих условиям гладкости. Типичный пример функция модуля

$$y = |x|$$

Поэтому мы ограничимся достаточно гладкими функциями, для которых применимы необходимые условия экстремума. Сформулируем достаточные условия

$$\Phi'(x^*) = 0$$

$$\Phi''(x^*) > 0$$

В таком случае, в достаточно малой окрестности точки x^*, разложение функции в ряд Тейлора с центром точки x^* имеет вид

$$\Phi(x) = \Phi(x^*) + \Phi'(x^*)(x - x^*) + \frac{1}{2}\Phi''(x^*)(x - x^*)^2 + o(|x - x^*|^2)$$

Поскольку x^* это точка экстремума

$$\Phi(x) = \Phi(x^*) + \frac{1}{2}\Phi''(x^*)(x - x^*)^2 + o(|x - x^*|^2)$$

Поведение функции в окрестности точки x^* в главном по величине $(x - x_x^*)$

определяется слагаемым, содержащим вторую производную. Для невырожденного экстремума $\Phi''(x^*) > 0$

$$\Phi(x) > \Phi(x^*), \quad x \neq x^*$$

Начнем с рассмотрения методов, которые не требуют информации о поведении производной функции, а только лишь о характере ее гладкости, так называемые методы нулевого порядка.

Итак, пусть точка единственного локального минимума x^* локализована на отрезке $[a, b]$. Функции, имеющие единственный локальный минимум на некотором множестве, называют унимодальными на этом множестве. Мы не знаем конкретное положение точки x^*, поэтому возникает вопрос о локализации положении функции на меньшем чем $[a, b]$ отрезке. То есть, провести итерационный процесс построения системы стягивающихся отрезков содержащих точку x^*.

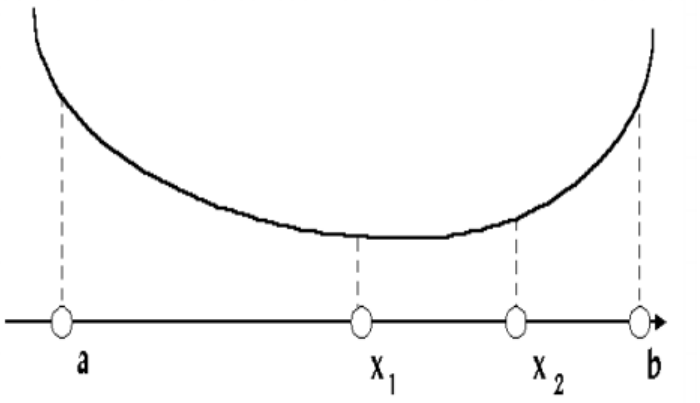

Рис. 12.1 – локализация корня

Изобразим, без ограничения общности, функцию на отрезке $[a, b]$ (рис. 12.1). Предположим, что у нас на отрезке находятся дополнительные точки x_1 и x_2 и наименьшее значение функции среди значении $f(a), f(b), f(x_1), f(x_2)$ достигается в точке x_1. Тогда из получившихся трех интервалов на $[a, b]$ мы можем отбросить интервал $[x_2, b]$. Таким образом, владея тремя точками и добавляя четвертую, можно получить систему вложенных отрезков, дальше сравнивая значения функции в

этих точках. Формализация этой процедуры может быть выполнена по-разному. Мы рассмотрим метод золотого сечения.

Будем говорить, что точка C делит отрезок $[a, b]$ по правилу золотого сечения, если выполнена следующая пропорция

$$\begin{cases} \frac{L}{V} = \frac{V}{U} = \tau > 1 \\ L = V + U \end{cases}$$

При этом нам неважно с какого конца откладывать отрезок U. Пропорция в которой делится отрезок точкой C и точкой D по-прежнему пропорции золотого сечения.

L-большая доля, U-меньшая доля.

Из этой системы получим параметр τ

$$\frac{L}{V} = 1 + \frac{U}{V}$$

$$\tau = 1 + \frac{1}{\tau}$$

$$\tau^2 - \tau - 1 = 0$$

$$\tau = \frac{1 + \sqrt{5}}{2}$$

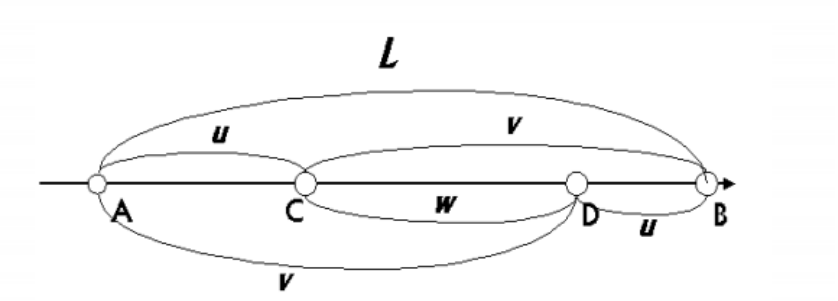

Рис. 12.2 – метод золотого сечения

Далее отложим на отрезке длины V отрезок длины U и напишем похожие соотношения для W и U

$$\begin{cases} \frac{V}{U} = \frac{U}{W} = \tau > 1 \\ V = W + U \end{cases}$$

Пусть длины L_k частичных отрезков, где локализовалась точка экстремума x^*, строились по правилу золотого сечения, тогда

$$\frac{L}{L_1} = \frac{L_1}{L_2} = \ldots = \frac{L_{n-1}}{L_n} = \tau$$

Перемножая эти пропорции, получим

$$L_n = \frac{L}{\tau^n} = q^n L$$

Это означает, что область локализации x^* соответствует алгоритму первого порядка точности.

12.2. Метод более высокого порядка

Использование информации о функции $f(x)$ и ее производных позволяет применить методы более высокого порядка точности. Опишем метод бисекции. Вычислим производную в середине отрезка $[x_{\text{лев}}, x_{\text{пр}}]$

$$x_0 = \frac{x_{\text{лев}} + x_{\text{пр}}}{2}; \quad f'(x)$$

Если в точке x_0 производная меньше нуля, то это означает что функция убывает, и тогда в качестве области локализации выберем новый отрезок $[x_{\text{лев}1} = x_0, x_{\text{пр}1} = x_{\text{пр}}]$. В случае когда производная больше нуля выбираем отрезок $[x_{\text{лев}1} = x_{\text{лев}}, x_{\text{пр}1} = x_0]$. Если же производная равна нулю и мы попали в стационарную точку, нужно разбираться с характером поведения функции в этой точке. Если нам дано, что функция унимодальна, тогда мы можем сказать сразу что попали в точку экстремума x^*. Тем самым повторяя эти действия для каждого выбранного сегмента, мы получим систему стягивающихся к точке x^* сегментов. Точка экстремума в этом случае на n-ом шаге будет локализована на отрезке длины

$$L_n = \frac{L}{2^{n+1}}; \qquad x \in [x_{\text{лев}n}, x_{\text{пр}n}]$$

Если воспользоваться характеристиками гладкости и формулами Тейлора можно применить более эффективные методы

$$f(x) = \underbrace{\Phi(x_k) + \frac{1}{2}\Phi''(x_k)(x - x_k)^2}_{\Psi_k(x)} + o(|x - x_k|^2)$$

Если ограничиться квадратичным приближением, то мы сможем заменить исходную задачу вычисления минимума функции $f(x)$ на вычисление минимума параболы $\Psi_k(x)$. Ветви этой параболы направлены вверх, потому что в окрестности невырожденного экстремума по достаточному условию $\Psi''(x^*) > 0$.

$$\Psi_k'(x) = \Phi'(x_k) + \Phi''(x_k)(x - x_k) = 0$$
$$x_{k+1} = x_k - \frac{\Phi'(x_k)}{\Phi''(x_k)}$$

Мы получили метод Ньютона поиска корня уравнения $\Phi'(x) = 0$, который обладает повышенной сходимостью ($|x_{k+1} - x^*| = O(|x_k - x^*|^2)$). Но при этом нам необходимо на каждой итерации вычислять производные $\Phi'(x_k)$ и $\Phi''(x_k)$ нашей функции.

Чтобы не прибегать к вычислению производных в явном виде можно по трем точкам x_{k-2}, x_{k-1}, x_k и соответствующим им значениям функции $\Phi_k(x)$ построить интерполяционный многочлен

$$\Phi(x) \approx N_2(x) = \Phi(x_k) + (x - x_k)\Phi(x_k, x_{k-1}) + (x - x_k)(x - x_{k-1})\Phi(x_k, x_{k-1}, x_{k-2})$$

Мы получаем параболу. В качестве очередного приближения x_{k+1} выбирают точку минимума построенной параболы

$$N_2'(x) = \Phi(x, x_{k-1}) + (2x - (x_k + x_{k-1}))\Phi(x_k, x_{k-1}, x_{k-2}) = 0$$

тогда

$$x_{k+1} = \frac{x_k + x_{k-1}}{2} - \frac{1}{2}\frac{\Phi(x_k, x_{k-1})}{\Phi(x_k, x_{k-1}, x_{k-2})}$$

*Замечание.*В методе Ньютона и в методе парабол обязательна проверка условия

$$\Phi(x_{k+1}) \leq \Phi(x_k)$$

Сходимость метода парабол выше чем линейная, но не квадратичная

$$(|x_{k+1} - x^*| = O(|x_k - x^*|^\alpha))$$

Задание. Найти коэффициент α.

12.3. Минимизация функции многих переменных

Рассмотрим задачу о построении локального экстремума функции многих переменных. Теперь наша целевая функции зависит от нескольких величин

$$\Phi(x_1, \ldots, x_m); \quad \vec{x} \in R^m; \quad \vec{x^*} = x^* \in loc\min_{R^m} \Phi(\vec{x})$$

Начнем с разложения функции по формуле Тейлора. Приращение функции запишем следующим образом

$$\Phi(\vec{x} + \vec{\triangle x}) = \Phi(\vec{x} + h\vec{p})$$

где $\vec{p}$ связан с направлением по которому мы движемся, h - шаг по этому направлению. В большинстве случаев $||\vec{p}|| = 1$ вектор коллинеарный вектору $\triangle x$ и тогда $h = ||\triangle x||$. Часто мы будем писать $h\vec{p} = \vec{p}$. Тогда

$$\Phi(\vec{x} + h\vec{p}) = d\Phi(\vec{x}) + \frac{1}{2!}d^2\Phi(\vec{x}) + o(||\triangle x||^2) =$$

$$= \Phi(\vec{x}) + \sum_{i=1}^{n}\frac{\partial\Phi}{\partial x_i}\vec{x}\triangle x_i + \frac{1}{2!}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial^2\Phi}{\partial x_i \partial x_j}\vec{x}\triangle x_i\triangle x_j + o(||\triangle x||^2) =$$

$$= \Phi(\vec{x}) + h(grad\Phi(x), \vec{p}) + \frac{h^2}{2}(G\vec{p}, \vec{p}) + o(h^2)$$

Где G - гессиан, матрица вторых производных

$$G = \left\| \frac{\partial^2\Phi}{\partial x_i \partial x_j}\vec{x} \right\|, \qquad G^T = G$$

Запишем необходимые условия экстремума. Первое условие это стационарность точки экстремума $\vec{x^*}$

$$||grad\Phi(\vec{x^*})|| = 0 \Leftrightarrow \frac{\partial\Phi}{\partial x_k}(\vec{x^*}) = 0$$

Второе условие это неотрицательность квадратичной формы

$$G \geq 0; \quad (G\vec{p}, \vec{p}) \geq 0$$

Запишем достаточные условия экстремума. Первое условие – это стационарность точки экстремума $\vec{x^*}$

$$||grad\Phi(\vec{x^*})|| = 0 \Leftrightarrow \frac{\partial \Phi}{\partial x_k}(\vec{x^*}) = 0$$

Второе условие это положительная определенность квадратичной формы

$$G > 0; \quad (G\vec{p}, \vec{p}) > 0; \quad p \neq 0$$

Запишем разложение Тейлора для точки экстремума обозначив $\triangle x = \vec{x^*} + \vec{p}$

$$\Phi(x) = \Phi(\vec{x^*}) + \underbrace{(grad\Phi(x^*), \vec{p})}_{\equiv 0} + \underbrace{\frac{1}{2!}(G(x^*)\vec{p}, \vec{p})}_{>0, p\neq 0} + o(||p||^2)$$

В главном по величине $||p||$ порядке поведение функции в окрестности точки x^* определяется поведением квадратичной функции. Для невырожденного экстремума $G > 0, \forall p \neq 0$

$$\Phi(x) > \Phi(x^*), \quad x \neq x^*$$

12.4. Квадратичная функция аргумента $\vec{x}$

Часто вместо того, чтобы рассматривать задачу минимизации функции $\Phi(\vec{x})$, будем рассматривать задачу о минимизации ее квадратичного приближения в окрестности некоторой точки

$$\Psi(\vec{x}) = \frac{1}{2}(A\vec{x}, \vec{x}) + (\vec{b}, \vec{c}) + c$$

С симметричной, невырожденной матрицей $A = A^T, detA \neq 0$. Найдем градиент $grad\Psi(\vec{x})$ и гессиан $G = hess\Psi(\vec{x})$ функции

$$\Psi(x + h\vec{p}) = \frac{1}{2}(A(\vec{x} + h\vec{p}), \vec{x} + h\vec{p}) + (\vec{b}, \vec{x} + h\vec{p}) + c =$$

$$= \left[\frac{1}{2}(Ax, x) + (b, x) + c\right] + h\left\{\frac{1}{2}(A\vec{x}, \vec{p}) + \frac{1}{2}(A\vec{p}, \vec{x}) + (\vec{b}, \vec{p})\right\} + \frac{h^2}{2}(A\vec{p}, \vec{p}) = \quad (47)$$

$$= \Psi(\vec{x}) + h(A\vec{x} + \vec{b}, \vec{p}) + \frac{h^2}{2}(A\vec{p}, \vec{p})$$

$$A\vec{x} + \vec{b} = grad\Psi, \quad A = hess\Psi(\vec{x})$$

В квадратных скобках слагаемые первой группы 0-ого порядка по h, в фигурных слагаемые второй группы 1-го порядка по h, без скобок слагаемые 2-го порядка по h.

В точке возможного экстремума градиент функции равен нулю, поэтому мы должны решать СЛАУ

$$A\vec{x} + \vec{b} = 0 \quad \Rightarrow \quad \vec{x^*} = A^{-1}b$$

Напишем разложение (47) в окрестности точки экстремума $\vec{x^*}$, учитывая что в случае симметричной матрицы у нее существует ортонормированный базис из собственных векторов $\{e_i\}_{i=\overline{1,m}}$, отвечающих собственным значениям $\{\lambda_i\}$

$$\Psi(\vec{x^*}+h\vec{p}) = \Psi(\vec{x^*})+\frac{h^2}{2}(A\vec{p},\vec{p}) = \Psi(\vec{x^*})+\frac{h^2}{2}\left(\sum_{i=1}^{m}\alpha_i\vec{e}_i, \sum_{k=1}^{m}\alpha_k\vec{e}_k\right) = \Psi(\vec{x^*})+\frac{h^2}{2}\sum_{k=1}^{m}\lambda_k\alpha_k^2$$

Таким образом, характер изменения $\Psi(\vec{x})$ при движении вдоль x_k полностью определяется знаком собственных значений $\{\lambda_k\}$.

12.5. Рельеф поверхности функции $\Psi(x)$. Линии уровня

Поверхность уровня. Для функции многих переменных, множество точек, где функция $\Psi(x_1,\ldots,x_n)$ принимает одинаковые значения

$$\Psi(x_1,\ldots,x_n) = \Psi(\overset{(0)}{x_1},\ldots,\overset{(0)}{x_n}) = C_0$$

Линия уровня. Для функции одной переменной множество точек, где

$$\Psi(x,y) = \Psi(x_0,y_0) = C_0$$

Если матрица A, кроме всего прочего, положительно определенная тогда все $\lambda_k > 0$ и поверхности уровня функции $\Psi(\vec{x})$ представляют собой множество вложенных m-мерных эллипсоидов или для функции одной переменной вложенных эллипсов, оси которых ориентированы по собственным векторам матрицы A. Нарисуем рельеф поверхности функции $\Psi(x)$ в окрестности экстремума.

Рис. 12.3 – а)-котловинный рельеф б)овражный рельеф

1) *Котловинный рельеф* - линии уровня представляют собой концентрические эллипсы. Чем меньше соответствующая константа, чем ближе мы к точке x^*.

2) *Овражный рельеф.* Если функция кусочно гладкая, то в линиях уровня наблюдается излом вектора касательной. Характеристики многих алгоритмов затруднены при наличии истинных оврагов.

Чаще приходится сталкиваться с разрешимыми оврагами когда поверхность уровня имеет котловинный рельеф

$$\Psi(x, y) = 100(y - x^2)^2 + (1 - x)^2$$

Видно, что точка с координатами (1,1) является глобальным минимумом. В рельефе функции выделяется серповидный овраг, дном которого является линия $y = x^2$(с дополнительными слагаемыми).

12.6. Спуск по координатам.

Перейдем к методам минимизации функции и в качестве первого рассмотрим метод покоординатного (циклического) спуска.

Опишем процедуру построения минимизирующей последовательности $\{M_k\}$ для целевой функции $\Phi(x_1, \ldots, x_k)$, для которой выполнено $\Phi(M_{k+1}) \leq \Phi(M_k)$. На каждом шаге будем фиксировать все координаты кроме выделенной, тем самым значения нашей функции будут зависеть только от одной переменной. Пусть на начальном этапе у нас была точка $\overset{(0)}{M}(\overset{(0)}{x_1}, \overset{(0)}{x_2}, \ldots, \overset{(0)}{x_m})$, фиксируя первую координату проведем спуск в точку $\overset{(1)}{M}(\overset{(1)}{x_1}, \overset{(0)}{x_2}, \ldots, \overset{(0)}{x_m})$ (индексом k будем обозначать количество спусков). Предположим, что какое-то количество циклов спуска уже выполнено и k переменных получили новые значения $\overset{(k)}{M}(\overset{(1)}{x_1}, \ldots, \overset{(1)}{x_k}, \overset{(0)}{x_{k+1}}, \ldots, \overset{(0)}{x_m})$. Опишем как перейти в точку $\overset{(k+1)}{M}$. Фиксируем все координаты кроме $k+1$-ой

$$\Phi(\overset{(1)}{x_1}, \ldots, \overset{(1)}{x_k}, x_{k+1}, \overset{(0)}{x_{k+2}}, \ldots, \overset{(0)}{x_m}) \equiv \varphi_{k+1}(x_{k+1})$$

$$\varphi(\underbrace{x_{k+1} + h_{k+1}\vec{e}_{k+1}}_{x_{k+1}}) = \min_h \varphi(\vec{x}_{k+1} + h\vec{e}_{k+1})$$

Решая задачу одномерной минимизации по параметру h, мы определяем минимум по направлению $\vec{e}_{k+1}$ и строим h_{k+1}. Тем самым мы приходим в точку

$$\overset{(k+1)}{M}(\overset{(1)}{x_1}, \ldots, \overset{(1)}{x_{k+1}}, \overset{(0)}{x_{k+2}}, \ldots, \overset{(0)}{x_m})$$

13. Лекция 13. Методы минимизации

13.1. Метод покоординатного спуска. Продолжение

Применим метод покоординатного спуска к квадратичной функции

$$\Psi(\vec{x}) = \frac{1}{2}(A\vec{x}, \vec{x}) + (\vec{b}, \vec{c}) + c$$

С SPD-матрицей (симметричной, положительно определенной). Напомним, что нам надо перейти из точки $\overset{(k)}{M}$ в точку $\overset{(k+1)}{M}$. Введем некоторые обозначения

$$\overset{(k)}{M}(\underbrace{\overset{(1)}{x_1}, \ldots, \overset{(1)}{x_k}, \overset{(0)}{x_{k+1}}, \ldots, \overset{(0)}{x_m}}_{\vec{x}_k}) \to \overset{(k+1)}{M}(\underbrace{\overset{(1)}{x_1}, \ldots, \overset{(1)}{x_{k+1}}, \overset{(0)}{x_{k+2}}, \ldots, \overset{(0)}{x_m}}_{\vec{x}_{k+1}})$$

Тогда

$$\vec{x}_{k+1} = \vec{x}_k + h_{k+1}\vec{e}_{k+1} \tag{48}$$

параметр h_{k+1} мы ищем из соображений минимизации функции при изменении только параметра h

$$\varphi_{k+1}(h) = \Psi(\vec{x}_k + h_{k+1}\vec{e}_{k+1}) = \Psi(x_k) + h(Ax_k + b, e_{k+1}) + \frac{h^2}{2}(Ae_{k+1}, e_{k+1}) = 0$$

То есть это одномерная задача, функция Ψ зависит только от параметра h и является параболой с ветвями направленными вверх, нам нужно найти ее минимум (вершину), найдя производную по h и приравняв ее к нулю

$$h_{k+1} = -\frac{Ax + b, e_{k+1}}{Ae_{k+1}, e_{k+1}}$$

Подставим найденной h_{k+1} в выражение (48)

$$\vec{x}_k - \frac{Ax + b, e_{k+1}}{Ae_{k+1}, e_{k+1}}\vec{e}_{k+1}$$

Нас будет интересовать только $k+1$-ая координата в этой записи, запишем ее, учитывая что в базисном векторе $\vec{e}_{k+1}$ все координаты нулевые, кроме $k+1$-ой на месте которой стоит единица

$$\overset{(1)}{x_{k+1}} = \overset{(0)}{x_{k+1}} - \frac{1}{a_{k+1,k+1}}\left(\sum_{j=1}^{k} a_{k+1,j}\overset{(1)}{x_j} + \sum_{j=k+1}^{m} a_{k+1,j}\overset{(0)}{x_j} + b_{k+1}\right) =$$

$$= -\frac{1}{a_{k+1,k+1}}\left(\sum_{j=1}^{k} a_{k+1,j}\overset{(1)}{x_j} + \sum_{j=k+2}^{m} a_{k+1,j}\overset{(0)}{x_j} + b_{k+1}\right) \Rightarrow$$

Перенесем все слагаемые в правую часть домножив на $a_{k+1,k+1}$

$$\sum_{j=1}^{k+1} a_{k+1,j}\overset{(1)}{x_j} + \sum_{j=k+2}^{m} a_{k+1,j}\overset{(0)}{x_j} + b_{k+1} = 0$$

Тем самым для системы уравнений $Ax + b = 0$ использован итерационный метод Зейделя, итерационная матрица и итерационный параметр которого

$$B = D + A_L; \qquad \tau = 1$$

Для SPD-матрицы было доказано, что этот метод сходится. Тем самым метод циклического покоординатного спуска примененный к задаче минимизации квадратичной функции с симметрично положительно определенной матрицей эквивалентен к методу Зейделя и $x_k \to x^*$.

13.2. Метод покоординатного спуска в общем случае

Посмотрим, что можно сделать в общем случае функции двух переменных $\Phi(x, y)$. Для этого сформулируем теорему.

Теорема 15.

Пусть множество уровня дважды дифференцируемой функции $\Phi(x, y)$

$$\{(x, y)\} : \quad \Phi(x, y) \le \Phi(x_0, y_0) = \Phi_0$$

представляет собой замкнутую ограниченную область D и всюду в ней выполнено

$$\Phi_{xx} \ge a > 0; \quad \Phi_{yy} \ge b > 0; \quad |\Phi_{xy}| \le \rho; \quad ab - c^2 > 0$$

тогда метод покоординатного спуска для $\forall x_0 \in D$

$$\vec{x}_k \to x^*$$

или что то же самое

$$M_k \to M^*; \ \ \forall M_0 = M_0(x, y) \in D$$

Проиллюстрируем, как выполняются циклы спуска.

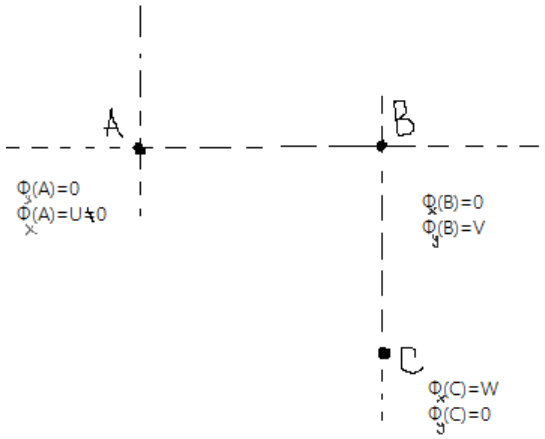

Рис. 13.4 – Схема спуска покоординатам

Первый шаг из точки A в точку B

$$\begin{cases} U = |\Phi_x(B) - \Phi_x(A)| = \Phi_{xx}(\xi)|x_B - x_A| \geq a\rho(A,B) \\ V = |\Phi_y(B) - \Phi_y(A)| = \Phi_{xy}(\xi)|x_B - x_A| \leq c\rho(A,B) \end{cases} \Rightarrow cU \geq aV$$

Второй шаг из точки B в точку C

$$\begin{cases} W = |\Phi_x(C) - \Phi_x(B)| = \Phi_{xy}(\eta)|y_C - y_B| \leq c\rho(C,B) \\ V = |\Phi_y(C) - \Phi_y(B)| = \Phi_{yy}(\overline{\eta})|y_C - y_B| \geq b\rho(C,B) \end{cases} \Rightarrow bW \leq cV$$

Из двух этих систем следует

$$W \leq \frac{c}{b}V \leq \frac{c^2}{ab}U$$

Так как ab больше чем c^2, тогда знаменатель нашей геометрической прогрессии меньше единицы

$$\frac{c^2}{ab} = q \quad 0$$

За один цикл спуска модуль компоненты $|\Phi(x,y)|$ уменьшился на величину q

$$\left| \overset{(n+1)}{\Phi_x} \right| \leq q \left| \overset{(n)}{\Phi_x} \right|$$

$$\left| \overset{(n+1)}{\Phi_y} \right| \leq q \left| \overset{(n)}{\Phi_y} \right|$$

Компоненты градиента убывают как геометрическая прогрессия. Тем самым на нашей траектории $\{M_n\}$, которая определяются циклами спуска, координаты градиента стремятся к нулю

$$\Phi_x(m_n) \to \infty, \quad \Phi_y(m_n), \to \infty \quad n \to \infty$$

Пользуясь тем, что область D находилась внутри множества уровня и эта область замкнута и ограничена(компактна), то есть из любой бесконечной последовательности мы можем выделить сходящуюся подпоследовательность. Мы будем считать этой подпоследовательностью построенную нами $\{M_n\}$. Отсюда вывод, что наша последовательность циклического координатного спуска сходится к M^*

$$\exists M^* = \lim_{n\to\infty} M_n$$

Для того чтобы наглядно в этом убедиться напишем разложение компонент градиента в точке M до первого порядке

$$\Phi_x(x,y) = \Phi_x(M^*) + \Phi_{xx}(M^*) \triangle x + \Phi_{xy}(M^*) \triangle y + \cancel{\cdots}$$
$$\Phi_y(x,y) = \Phi_y(M^*) + \Phi_{yx}(M^*) \triangle x + \Phi_{yy}(M^*) \triangle y + \cancel{\cdots}$$

Если подставит в эти выражения точки нашей траектории, то

$$\triangle x = x_k - x^*; \quad \triangle y = y_k - y^*; \quad \Phi_x(M^*) = \Phi_y(M^*) = 0$$

Мы получаем СЛАУ, которая однозначна разрешима, поскольку определить гессиана больше нуля.

13.3. Метод наискорейшего спуска

Рассмотрим разложение целевой функции $\Phi(\vec{x})$ в окрестности точки $\vec{x}_k$

$$\Phi(\vec{x}) = \Phi(\vec{x_k}) + (\underbrace{grad\Phi(\vec{x_k})}_{\vec{g}_k}, \vec{p}) + o(||\vec{p}||)$$

Попытаемся понять, как нам выбрать направление, чтобы обеспечить наискорейшее изменение функции $\Phi(\vec{x})$. Для нас понятно, что направление нужно выбирать по антиградиенту, но мы сформулируем задачу поиска направления как задачу минимизации

$$\begin{cases} min(\vec{g_k}, \vec{p}) \\ ||\vec{p}||_A^2 = 1 = (A\vec{p}, \vec{p}) \end{cases}$$

$||\vec{p}||_A^2$-энергитическая норма, A - SPD-матрица. Чтобы решить эту задачу, построим функцию Лагранжа

$$L = (\vec{g_k}, \vec{p}) + \lambda((A\vec{p}, \vec{p}) - 1)$$

Легко вычислить градиент получившейся функции

$$\vec{g}_k + 2\lambda A\vec{p} = 0$$

Отсюда мы можем понять, что такое направление $\vec{p}$

$$\vec{p} = -\frac{1}{2\lambda}A^{-1}\vec{g}_k$$

Отсюда делаем вывод, что направление $\vec{p}$ (при его нормировке) можно выбирать как $-A^{-1}\vec{g}_k$. Для случая, когда $A = E$ – это направление антиградиента $\vec{p} = -\vec{g}_k$, что и приводит нас к методу наискорейшего спуска

$$\begin{cases} \vec{x}_{k+1} = \vec{x}_k + h_k\vec{p}_k \\ h_k : \quad \varphi(h_k) = \Phi(\vec{x}_k + h\vec{g}_k) \end{cases}$$

Рассмотрим применение метода наискорейшего спуска к минимизации квадратичной функции.

$$\Psi(\vec{x}) = \frac{1}{2}(A\vec{x}, \vec{x}) + (\vec{b}, \vec{c}) + c$$

Градиент $\vec{g} = Ax + b$, тогда $p_k = -g_k = a(Ax + b)$

Необходимо изучить эту функцию на векторе $\vec{x}_k + h\vec{p}_k$ и минимизировать ее по h. Иногда мы будем опускать векторы, помня что мы работаем с функцией многих переменных.

$$\Psi(x_k + hp_k) = \Psi(x_k) + h(Ax_k + b, p_k) + \frac{h^2}{2}(Ap_k, p_k)$$

$$h_{k+1} = -\frac{(g_k, p_k)}{(Ap_k, p_k)} = \frac{(g_k, p_k)}{(Ag_k, g_k)}$$

$$\vec{x}_{k+1} = \vec{x}_k - \frac{||g_k||^2}{||g_k||_A^2}, \vec{g}_k$$

Просьба дать интерпретацию этому методу как итерационному методу решения алгебраической системы.

Один из основных недостатков метода наискорейшего спуска состоит в том, что направление антиградиента является локальным.

13.4. Методы второго порядка. Метод сопряженных градиентов

Все предыдущие методы для минимизации квадратичной функции требовали бесконечное число итераций. Поэтому рассмотрим более точные методы. Запишем разложение функции $\Phi(x)$ в окрестности точки x_k

$$\Phi(x_k + p) = \underbrace{\Phi(x_k) + (grad\Phi(x_K), p) + \frac{1}{2}(G(x_p)p, p)}_{\Psi_k(p)} + o(||p||^2)$$

Будем минимизировать функцию $\Psi(p)$. Продифференцируем функцию по параметру p

$$\frac{\partial\Psi(p)}{\partial p} = \underbrace{grad\Psi_k}_{\vec{g}_k} G_k\vec{p} = 0$$

$$\vec{p} = -G_k^{-1}\vec{g}_k$$

Тем самым за одну итерацию(метод Ньютона) можно выбрать квадратичное приближение для функции $\Psi(x)$ и минимизировать функцию $\Psi(p)$. Мы получаем метод не ниже второго порядка сходимости, но при этом нам необходимо обратить матрицу G_k, что не всегда удобно.

Перейдем к методу сопряженных градиентов. Введем определения.

Пусть есть SPD-матрица A, тогда два вектора называются сопряженными относительно этой матрицы, если

$$(x, y)_A = (x, Ay) = (Ax, y) = 0$$

$(x, y)_A$ - энергетическое скалярное сопряжение. Ортогональные в смысле энергетического скалярного произведения векторы называются A-сопряженными векторами. Если у нас есть набор A-сопряженных векторов $x_1, \ldots, x_k (k \leq m)$, то такие векторы являются линейно независимыми. Действительно, если предположить что один из векторов, например x_1, является линейной комбинацией остальных, тогда

$$||x_1||_A^2 = (x_1, A\sum_{p=2}^{k}\alpha_p x_p) = \sum_{p=2}^{k}\alpha_p(x_1, Ax_p) = 0$$

Получаем противоречие. Тем самым m векторов будут образовывать ортонормированный базис $\{x_i\}$ в нашем пространстве.

$$(x_k, x_l)_A = \delta_{kl}$$

Разложим направление $\vec{p}$ по этому базису и рассмотрим квадратичную функцию на этом направлении

$$\Psi(x_k + p) = \Psi(x_k) + (Ax_k + b, p) + \frac{1}{2}(Ap, p) =$$

$$= \Psi(x_k) + \left(Ax + b, \sum_i \alpha_i x_i\right) + \frac{1}{2}\left(\sum_{i=1}^{m} \alpha_i Ax_i, \sum_{j=1}^{m} \alpha_j Ax_j\right)$$

$$= \Psi(x_k) + \sum_{i=0}^{m} \alpha_i (grad\Psi_k, x_i) + \sum_{i=1}^{m} \frac{\alpha_i^2}{2}$$

Изменение квадратичной функции в окрестности точки x разбивается на m независимых слагаемых каждое из которых связанно лишь с одним сопряженным вектором. В итоге, для того чтобы минимизировать функцию, нужно минимизировать каждое из слагаемых. При этом $grad\Psi(x_k) = b$(фиксирован), матрица $G(x_k)$ - одна и та же при движении по всем направлениям процедуры сопряжения градиентов. Функцию $\Psi_k(p)$ переобозначим в функцию $\Psi(x)$. Опишем процедуру построения одного цикла минимизации, содержащего n шагов и точно минимизирующего функцию общего вида $\Psi_k(\vec{p})$

$$\text{Цикл движения} \quad M_k \equiv \overset{(1)}{M_0} \underset{(p_1)}{\longrightarrow} \overset{(2)}{M_0} \underset{(p_2)}{\longrightarrow} \overset{(3)}{M_0} \longrightarrow \ldots \underset{(p_s)}{\longrightarrow} \overset{(s+1)}{M_0} \underset{(p_{s+1})}{\longrightarrow} \overset{(s+2)}{M_0} \ldots \underset{(p_n)}{\longrightarrow} M_{k+1}$$

Нам необходимо построить G-сопряженные направления, минимизация по каждому из них будет приводить нас не более чем за n шагов в точку минимума квадратичной функции. Точка M_{k+1} будет использоваться как следующий шаг. Если считать что p направления уже построены, то следующие направление, кроме первого, строятся по одним формулам

$$p_{s+1} = -g_{s+1} + \alpha_s p_s$$

$$\alpha_s = \frac{g_{s+1}^2}{g_s^2}$$

$$x_{s+2} = x_{s+1} + h_{s+1} p_{s+1}$$

Где h выбирается из минимизации квадратичной функции

$$\Psi(x_{s+1} + hp_{s+1})$$

Для первого направления

$$p_1 = -g_1$$

Утверждается, что совокупность построенных направлений G-сопряженные.

14. Лекция 14. Минимизация функционала

Пусть любому $y \in Y$ поставлено в соответствие число, тогда говорят что на множестве функций Y задан функционал $\Phi[x] \in R$. Ставиться задача поиска такого y^*, что значение функционала на функции y^* доставляет минимум функционала на всем множестве Y

$$\Phi^* = \Phi[y^*] = \inf_Y \Phi[y]$$

Таким образом, в будущем у нас будет строиться две последовательности. Последовательность функционалов

$$\Phi_n \to \Phi^*$$

И последовательность аргументов

$$y_n \to y^*, \quad y^* \in arg\min_Y \Phi[y]$$

Поэтому различают задачи минимизации функционала и задачу минимизации аргументов. Как правило, Y будет некоторым гильбертовом пространством. Построим его как множество n-параметрического семейства функций

$$Y_n = \{y_n(\vec{x}, a_1, \dots, a_n)\}$$

На этом множестве значение функционала превращается в функцию от параметров $a_1, \dots, a_n$

$$\Phi(y_n(x, \vec{a})) = F(a_1, \dots, a_n)$$

Тогда задача нахождения a^* ставится как задача минимизации функции $F(a_1, \dots, a_n)$. Задача на безусловный экстремум в R^n

$$a^*: \quad m\min_{a \in R^n} F(a_1, \dots, a_n) = F^*(\vec{a}^*)$$

Отвечающую этому набору параметров функцию будем обозначать

$$y_n^*(\vec{x}, \vec{a}^*); \quad \Phi_n^* = \Phi[y_n^*(x, \vec{a}^*)]$$

14.1. Сходимость последовательности значений функционала и последовательности аргументов

Покажем, что есть возможность построить последовательность $\{\Phi_n^*\}$, сходящуюся к Φ^*. Для этого построим систему вложенных классов (нижний индекс указывает на кол-во параметров).

$$Y_1 \subset Y_2 \subset \dots \subset Y_n \subset \dots$$

На каждом из этих классов мы решаем задачу о минимизации соответствующего функционала на n-параметрическом семействе функций

$$\Phi_1^* \geq \Phi_2^* \geq \ldots \geq \Phi_n^* \geq \ldots \geq \Phi_1^*$$

Таким образом, последовательность $\{\Phi_n^*\}$ является невозрастающей и ограниченной снизу, а значит имеет предел

$$\{\Phi_n^*\} \to \overline{\Phi}^* \geq \Phi^*$$

Сформулируем достаточные условия сходимости последовательности $\{\Phi_n^*\}$

1)$\Phi[y]$ - непрерывный функционал.

Функционал является непрерывным, если на функции $y_0(x)$ если для $\forall \varepsilon > 0$ можно указать $\delta(\varepsilon) > 0$, такое что

$$||y(x) - y_0(x)|| < \delta \Rightarrow \left|\Phi[y] - \Phi[y_0]\right| < \varepsilon$$

Если это условие выполнено для всех функций y, то она является непрерывным на множестве Y.

2) Необходима предельная полнота семейства функций $\{Y_n\}$, $n = \overline{1, \infty}$.

Мы говорим, что семейство функций обладает свойством полноты, если для $\forall y \in Y$ и для $\forall \varepsilon$ существует такой номер N, что для любого $n > N$ найдется функция $\tilde{y}_n(x, \vec{a})$, такая что

$$||\tilde{y}_n(x, a) - y(x)|| < \varepsilon$$

Теорема 16

Пусть выполнены достаточные условия, тогда последовательность $\{y_n^*(x, \vec{a}^*)\}$ является минимизирующей

$$\Phi_n^*[y_n^*] \to \Phi^*, \quad n \to \infty$$

Замечание. Отсюда еще не следует, что сама последовательность $\{y_n^*\}$ сходится к y^*. Эта теорема доказывает лишь сходимость последовательности значений функционала, принимаемых на последовательности $\{y_n^*\}$.

Доказательство

По $\delta(\varepsilon)$, которое фигурирует в первом достаточном условии, найдется номер N_0, такой что для любого $n > N_0$ в каждом из наших n-параметрических пространств $(n > N)$ найдется функция $\tilde{y}_n(x, \vec{a})$ такая что

$$\exists \tilde{y}_n(x, \vec{a})$$

$$||\tilde{y}_n(x, \vec{a}) - y^*|| < \delta$$

Раз это требование выполнено, то в силу непрерывности

$$\tilde{\Phi}_n - \Phi^* < \varepsilon$$

Если условие выполнено, тогда оно будет выполнено и для меньшего значения функционала

$$\tilde{\Phi}_n^* - \Phi^* < \varepsilon$$

Тем самым мы показали сходимость последовательности $\{\Phi_n^*\}$. Это и означает, что последовательность $\{y_n^*\}$ является минимизирующей.

В множестве функций $\tilde{y}_n(x, \vec{a})$, которые могли найтись при доказательстве теоремы 16, была и функция y^*, к которой сходится последовательность $\{y_n^*\}$. Но из этого еще не следует сходимость по аргументу, поэтому приступим к доказательству сходимости.

Теорема 17

Пусть выполнены условия теоремы 16 и существует такая δ-окрестность функции y^*, что

$$\Phi[y] - \Phi[y^*] \geq \alpha||y - y^*||^\beta, \quad \alpha, \beta > 0, \quad ||y - y^*|| < \delta$$

Тогда мы можем утверждать что

$$\alpha||y_n^* - y^*||^\beta \Phi_n^* - \Phi^* < \varepsilon \Rightarrow$$

$$||y_n^* - y^*|| \to 0, \quad n \to \infty$$

То есть

$$y_n^* \to y^*, \quad n \to \infty$$

14.2. Задачи минимизации функционала

Постановка задачи. Пусть положительный ($(Ax, x >; 0)$ при $x \neq 0$) оператор A с симметричной матрицей действует в гильбертовом пространстве $A : H \to H$. Будем считать что пространство H является энергетическим пространством оператора A. Решается задача

$$Ay = f \tag{49}$$

Покажем, что эта задача эквивалентна задаче минимизации квадратичного функционала $\Phi[y]$

$$\Phi[y^*] = \min_Y \Phi[y] \tag{50}$$

а) Запишем квадратичный функционал

$$\Phi[y] = \frac{1}{2}(Ay, y) - (f, y) \tag{51}$$

Любую функцию y можно представить как $y = y^* + h\delta y(x)$, тогда

$$\Phi[y] = \Phi[y^*] + h\delta y = \Phi[y^*] + h(Ay^* - f, \delta y) + \frac{h^2}{2}(A\delta y, \delta y)$$

Если y^* является решением задачи (49), тогда

$$(Ay^* - f, \delta y) \equiv 0$$

И тем самым для произвольной точки y

$$\Phi[y] \geq \Phi[y^*]$$

Это следует из положительности оператора A.
Значит y^* является решением задачи (50)

б)

$$\delta\Phi[y^*, \delta y] = \frac{d}{dh}\Phi[y^* + h\delta y]\Big|_{h=0} = 0 = (Ay^* - f, \delta y); \quad \forall \delta y \in Y$$

В том числе это верно для $\delta y = Ay^* - f$, а значит

$$||Ay^* - f|| = 0 \Leftrightarrow Ay^* = f$$

Это означает, что y^* является решением задачи (49). Тем самым мы видим что задачи (49) и (50) эквивалентны.

Поэтому в случае, когда мы имеем дело с положительной симметричной матрицей, в качестве приближенного метода решения предлагают минимизацию эквивалентного функционала $\Phi[y]$.

14.3. Метод Ритца

Если строить Y_n как линейную оболочку порожденную функциями $y_1(x), \ldots, y_n(x)$

$$Y_n = Lin(y_1(x), \ldots, y_n(x)), \quad \{y_i(x)\} - \text{базис в } Y$$

То есть

$$y_n(x, a) = \sum_{k=1}^{n} a_k y_k(x) = \sum_{k=1}^{n} a_k \varphi_k(x)$$

Тогда наш квадратичный функционал (51) записывается как

$$\Phi[y_n(x, a)] = F(a_1, \ldots, a_n) = \frac{1}{2}\sum_k \sum_l a_k a_l (A\varphi_k, \varphi_l) - \sum_k a_k (f, \varphi_k)$$

Квадратичная функция переменных $a_1, \ldots, a_n$. Запишем решение задачи минимизации для квадратичной функции

$$\frac{\partial F}{\partial a_p} = \sum_{m=1}^{n} (A\varphi_p, \varphi_m) - (f, \varphi_p) = 0$$

Получаем нормальную систему относительно коэффициентов a_k

$$\sum_{m=1}^{n} a_m (\varphi_m, A\varphi_p) = (f, \varphi_p)$$

Если система базисных функций полна в a-энергетической метрике, то мы можем утверждать, что последовательность $\{\Phi_n^*\}$ является минимизирующей, и

1)

$$\Phi_n^* \to \Phi^*$$

2)

$$||y_n^* - y^*||_A \to 0$$

15. Лекция 15. Разностные методы решения задач математической физики. Часть 1

Универсальным методом приближенного решения, применимым для широкого круга задач математической физики, является метод конечных разностей. Мы будем рассматривать следующие объекты. Точка $M(x, y, z)(x, y, z$ - имеют смысл пространственных координат) принадлежащая области G с границей ∂G, переменная, которая имеет смысл времени $t \in [t_0, T]$, как правило, мы будем полагать $t_0 = 0$. Также будем рассматривать расчетную область D:

$$D = G \times (t_0, T]$$

Граница Γ расчетной области

$$\Gamma = \partial G \times (t_0, T]$$

Дополнительные начальные условия в для точки t_0 ставятся в замкнутой области $\overline{G}$. Пока абстрагируемся от координат x, y, z, t и будем рассматривать задачи для переменных $x_1, \ldots, x_p$. Сформулируем исходную задачу

$$Au = f, \quad x \in D \tag{52}$$

$$Ru = g, \quad x \in \Gamma \tag{53}$$

Тем самым у нас есть операторное уравнение, которое выполняется во внутренних точках области D, и дополнительные условия для случаев, когда точка попадает на границу области Γ. Приближенное решение задачи (52,53) в методе конечных разностей строится с помощью разностной схемы. Для этого в расчетной области $\overline{D} = D \cup \Gamma$ нужно ввести расчетную сетку, состоящую из множества внутренних узлов ω_h и множества граничных узлов $\gamma_h : \Omega_h = \omega_h \cup \gamma_h = \{x_i\}_{i \in I}$. Задача (52) формулируется на сетке ω_h, а задача (53) на сетке γ_h. Мы пока абстрагируемся от смысла параметра "h" в соответствующих сетках, контролирующего как пространственные, так и временные размеры сетки. Рассмотрим сеточную функцию дискретного аргумента $y(x_i) = y_i = y_h$ на наборе точек $\{x_i\}$ и сформулируем для нее разностные задачи(или разностные схемы)

$$A_h y = \varphi_h \tag{54}$$

$$R_h y = \chi_h \tag{55}$$

Мы символически обозначили операторы в задачах выше теми же символами что и в задаче (52,53), но нужно понимать что это операторы, действующие на функцию не непрерывного переменного, а дискретного.

Приведем пример того, как некоторый оператор L_h действует на функцию y

$$L_h y = \sum_{k \in Ш_L(x)} \alpha_k y_k$$

$Ш_L(x)$- шаблон оператора L_h, который в общем случае может меняться в зависимости от того в окрестности какой точки мы рассматриваем разностную схему. В

большинстве случаев стараются чтобы шаблон носил регулярный характер и не зависел от выбранной точки.

Нам нужно ответить на вопросы как соотносится решение задач (52,53) с решением задач (54,55), как строится решение задач (54,55), что происходит с изменением параметра h и рассмотрением большего числа точек в интересующей нас расчетной области $\overline{D}$.

Введем понятие невязки разностной схемы. Если подставить решение разностной задачи y в нашу исходную задачу, то вряд ли нам получится добиться правильного равенства. Отсюда возникают невязки, первая

$$\Psi_n = \varphi_h - A_h u = (Au - f)_h - (A_h u - \varphi_h) = \varphi_h - A_h u$$

Величина $(Au - f)_h$ в точности равна нулю, поскольку решение u удовлетворяет (52,53). Мы переписали невязку в таком виде, потому что как правило решение u для нас в начале неизвестно. Поэтому мы будем требовать чтобы невязки обладали нужными нам свойствами для некоторого класса функций. Мы посмотрим это на примере аппроксимации. Вторая невязка связана с граничными условиями

$$\eta_n = \chi_h - R_h u = (Ru - g)_h - (R_h u - \chi_n)$$

15.1. Аппроксимация разностной схемы

Мы говорим, что разностная схема (54,55) аппроксимирует задачи (52,53) если

$$||\Psi_n|| \to 0 \quad ||\chi_n|| \to 0 \quad h \to 0$$

Удобно записать невязки в виде

$$\Psi_h = (Av - f)_h - (A_h v - \varphi_h); \quad \in V$$

Мы берем некоторый набор функций V, подразумевая что точное решение также принадлежит V. Далее подставим функцию v в (52) и получим невязку, так как функция v не является точным решением. Но величина этого несоответствия для тех схем, которые обладают аппроксимацией, согласована с подстановкой функции v в нашу разностную задачу, то есть разностная задача дает такое же по порядку величины несоответствие. Тем самым Ψ_h при измельчении h стремиться к нулю. То же самое и для невязки η_n.

Мы говорим что аппроксимация имеет k-ый порядок, если

$$||\Psi_h|| = O(h^*), \quad ||\eta_h|| = O(h^*)$$

При этом не всегда оказывается так, что аппроксимация одинаковой по разным переменным.

15.2. Устойчивость

Следующее требование для решения задачи (54,55) это устойчивость разностной схемы. Формально речь идет о непрерывной зависимости решений входных данных(правая часть разностной схемы). Охарактерезуем погрешность решения разностной задачи

$$||\delta y|| = ||\delta y^{(in)}|| + ||\delta y^{(app)}||$$

Изобразим типичный график зависимости погрешности сеточного решения от величины шага.

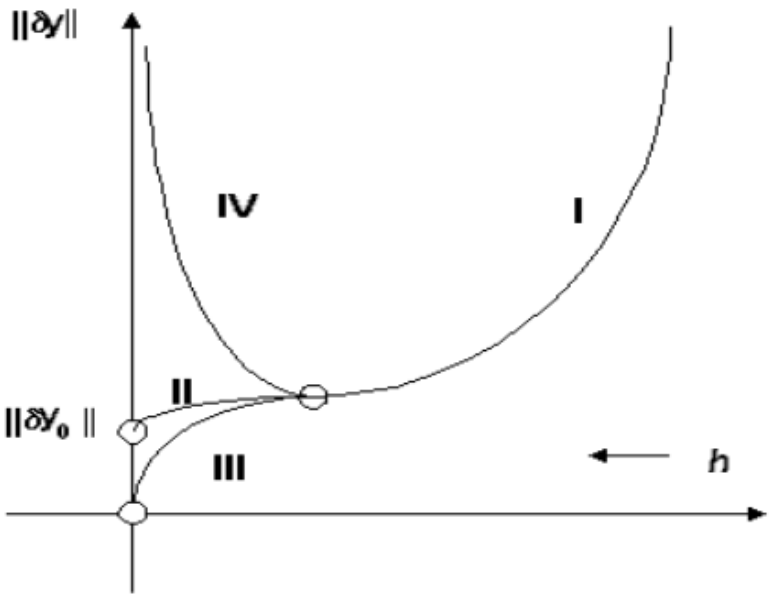

Рис. 15.1 – рост погрешностей

Траектория I. При уменьшении шага сначала погрешность всех схем убывает, так как существенно уменьшается погрешность аппроксимации. II—Для устойчивых схем погрешность сеточного решения будет стремиться к конечной величине связанной с ошибкой входных данных. Если при $h \to 0$ ошибка входных данных исчезает, то это случай III. То есть устойчивая схема в этом случае позволяет получить сколь угодно высокую точность расчета. Если же схема неустойчива (IV), то при $h \to 0$ погрешность δy_h возрастает до бесконечности (ибо растет объем неустойчивых вычислений).

Как правило, погрешности входных данных и аппроксимации имеют степенной характер зависимости от $h \Rightarrow h^a$, а неустойчивость приводит к возрастанию погрешности решения по экспоненциальному закону $\sim b^{\frac{a}{h}}$.

Дадим формальное определение устойчивости по входным данным φ и χ. Мы говорим, что схема устойчива если для любого $\varepsilon > 0$ существует $\delta(\varepsilon) > 0$ такое, что

$$\begin{array}{l} ||\varphi_1 - \varphi_2|| < \delta \\ ||\chi_1 - \chi_2|| < \delta \end{array} \quad \Rightarrow \quad ||y_1 - y_2|| < \varepsilon$$

Для линейных разностных схем можно переписать оценку погрешности как

$$||y_1 - y_2|| \le C_1||\varphi_1 - \varphi_2|| + C_2||\chi_1 - \chi_2||$$

Наша задача может иметь эволюционный тип, когда развитие по переменной t при достижении какого-то дискретного значения t^* определяется лишь состоянием системы в точке t^*, а не всей историей предыдущих шагов. Подобные задачи мы представляем как двуслойные разностные схемы.

15.3. Двуслойные разностные схемы

Пусть у нас есть сеточный оператор B_h и A_h действующие на y. Теперь наша сеточная функция будет зависеть от двух переменных, пробегающих дискретное множество точек

$$y = (x, t) \Rightarrow y(x_n, t_m) = y_n^m$$

После того как мы перейдем на следующий временной слой

$$y(x_n, t_{m+1}) \equiv \hat{y}$$

Тогда общий вид двуслойной разностной схемы

$$B_h \frac{\hat{y} - y}{\tau} + A_h y = \varphi_h$$

Опять же, переход на следующий слой будет осуществляться через сеточные значения y на предыдущем слое. Среди всех χ_h выделяют начальное условие $\chi_h\Big|_{t_0}$.

Разностная схема (54,55) называется *равномерно устойчивой* по начальным данным, если при постановке начальных данных на любом слое t ($t_0 \le t^* < t \le T$), она по ним устойчива и эта устойчивость не зависит от положение промежуточного слоя.

Для линейных разностных схем это означает, что существует $C > 0$ не зависящее от t^* и h и

$$||y_1(t) - y_2(t)||_{y_h} \le C||y_1(t^*) - y_2(t^*)||, \quad t_0 \le t^*$$

Из равномерной устойчивости следует обычная устойчивость.

Теорема 18. достаточный признак равномерной устойчивости

Пусть y_1 и y_2 решения разностной задачи с одинаковыми входными данными, но отвечающие различным начальным условиям. Тогда, если при переходе на следующий временной слой

$$||\hat{y}_1 - \hat{y}_2|| \le (1 + c\tau)||y_1 - y_2||)$$

наша схема равномерно устойчива

Доказательство

Предположим, что на каком-то слое t^* в нашем решении содержится погрешность δy, тогда на следующем временном слое погрешность $\delta\hat{y}$ удовлетворяет

$$||\delta\hat{y}|| \le (1 + c\tau)||\delta y|| \le (1 + c\tau)^{\frac{T-t^*}{\tau}}||\delta y|| \le (1 + c\tau)^{\frac{T-t_0}{\tau}}||\delta y_0|| \le A||\delta y_0||$$

Теорема 19. достаточный признак равномерной устойчивости по правой части

Если разностная схема (54,55) равномерно устойчива по начальным условиям и y_1, y_2, являющиеся решениями задачи $A_h y_{1,2} = \varphi_{1,2}$ с одними и теми же начальными условиями, совпали на каком-то временном слое и имеет место оценка

$$y_1(t_m) = y_2(t_m) \Rightarrow ||\hat{y}_1 - \hat{y}_2|| \le C\tau||\varphi_1 - \varphi_2||$$

тогда соответствующая разностная схема устойчива по правой части.

Доказательство

Для начала построим разностную схему (рис.15.2). Пусть у нас есть решения двух задач $A_h y = \varphi_h$ и $A_h\tilde{y} = \tilde{\varphi}_h$ отвечающие одному и тому же начальному условию $\chi_h\Big|_{t_0} y = \chi_h\Big|_{t_0} \tilde{y} = y_0$. Мы хотим охарактеризовать как отличаются y и $\tilde{y}$ на протяжении временного слоя. На временном слое рассматриваются дополнительные сеточные функции

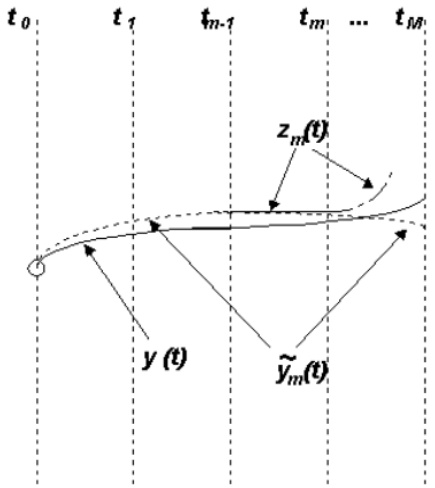

Рис. 15.2 – настоящее и приближенное решение

$$\begin{cases} z_m(t_{m-1}) = z_{m-1}(t_{m-1}) \\ A_h z + m = \left\{ \tilde{\varphi}, t_{m-1}\varphi, t > ; ; t_m \right. \\ y_0 = y(t_0) = \tilde{y}(t_0) = z_1(t_0) = z_0(t_0) \end{cases}$$

Тогда утверждается, что на каждом из слоев $t \in [t_{m-1}, t_m]$ решение возмущенной задачи $\tilde{y}(t)$ совпадает с соответствующей функцией $z_m(t)$, поскольку в точку t_{m-1} начальное условие принесено функцией z_{m-1}, удовлетворяющей возмущенному уравнению на соответствующем отрезке t. Аналогично на предыдущем слое и так далее, пока мы не попадем в начальную точку t. В точке $t = t_{m-1}$ и $\tilde{y}$ и z_1 имеет одно и то же начальное условие и на интервале (t_{m-1}, t_m) удовлетворяют возмущенной задаче. В силу единственности решения разностной задачи эти функции совпадают.

Далее, при $t \in (t_m, t_{m+1})$, функции $z_{m+1}(t)$ и $z_m(t)$ совпадают в точке t_m и удовлетворяют различным уравнениям, тогда

$$||z_m(t_{m+1}) - z_{m+1}(t_{m+1})|| \leq C\tau||\varphi - \tilde{\varphi}|| \to$$

В силу равномерной устойчивости по начальным данным мы можем написать, что

$$\Rightarrow ||z_{m+1}(t) - z_m(t)|| \leq C_1 C\tau||\varphi - \tilde{\varphi}||, \quad t^* \leq \forall t \leq t_m$$

Мы получаем

$$||z_{m+1}(t_m) - z_m(t_m)|| \leq A\tau||\varphi - \tilde{\varphi}||$$

Воспользуемся полученным неравенством и неравенством треугольника, чтобы посмотреть, что происходит с нашими решениями основной и возмущенной задачи

$$||\tilde{y}(t_m) - y(t_m)|| = ||z_m(t_m) - z_0(t_m)|| \leq ||z_m - z_{m-1}|| + ||z_{m-1} - z_{m-2} + \ldots + ||z_1 - z_0|| \leq$$
$$\leq MA\tau||\varphi - \tilde{\varphi}|| = A(T - t_0)||\varphi - \tilde{\varphi}||$$

Тем самым требования теоремы выполнены.

15.4. Сходимость разностной схемы

Схема называется сходящейся, если при h, стремящемся к нулю, погрешность решения также стремиться к нулю

$$||y-u||_h \to 0, \quad h \to 0$$

Так как u непрерывная функция, а y сеточная, мы предполагаем что функция u удобным образом спроецирована в то пространство где находится y.

Теорема 20.

Пусть u является решением задачи (52,53), а разностная схема (54,55) - корректна (решение существует,единственно,устойчиво) и аппроксимирует задачу (52,53), то

$$||y-u||_h \to 0, \quad h \to 0$$

Замечание. В жаргоне формулировка теоремы звучит так : "аппроксимация плюс устойчивость равно сходимость".

Доказательство

Запишем невязку для разностной схемы (54,55)

$$\begin{array}{l} \Psi_n = \varphi_h - A_h u = (Au-f)_h - (A_h u - \varphi_h) \\ \eta_n = \chi_h - R_h u = (Ru-g)_h - (R_h u - \chi_n) \end{array} \Leftrightarrow \begin{array}{c} A_h u = \varphi_h - \psi_h \\ R_h u = \chi_h - \eta_h \end{array} \tag{56}$$

То есть мы видим, что решение непрерывной задачи удовлетворяет разностной схеме возмущенной невязки. Из устойчивости решения

$$\begin{array}{l} ||\psi_h|| \le \delta(\varepsilon) \\ ||\eta_h|| \le \delta(\varepsilon) \end{array} \Rightarrow \quad ||y-u||_h < \varepsilon$$

Разностная схема обладает аппроксимацией, а это значит что $\exists h_0$, что для $\forall h < h_0$ выполнено условие

$$||\psi_h||, ||\eta_h|| < \varepsilon$$

Таким образом, мы показали, что для любого $\varepsilon > 0$ существует h_0, что при всех h

$$||y-u||_h < \varepsilon$$

А это и означает, что

$$||y-u||_h \to 0, \quad h \to 0$$

Тем самым теорема доказана.

Сходимость имеет порядок k, если

$$||y-u|| = O(h^*)$$

Замечание. Как правило разностная схема по различным переменным имеет различный порядок аппроксимации, например, невязка уравнения

$$||\psi_h|| = O(h^k + \tau^p), \quad h \to 0, \ \tau \to 0$$

Аналогично записывается невязка для $||\eta_h||$. Такая аппроксимация называется абсолютной, в отличии от условной аппроксимации в случае, когда

$$||\psi_h|| = O\left(h^k + \tau^p + \frac{\tau^t}{h^\delta}\right), \quad h \to 0,\ \tau \to 0,\ \frac{\tau^t}{h^\delta} \to 0$$

Аналогично для $||\eta_h||$.

Теорема 21.

Если выполнены условия теоремы 20 и разностная схема линейна, то сходимость не ниже порядка аппроксимации.

Доказательство

Запишем погрешность разностного решения

$$z_h = (y - u)_h$$

Для этого вычтем из уравнений (54,55) полученную разностную схему возмущенную невязками (56). Получим

$$A_h z = \psi_h$$
$$R_h z = \eta_h$$

Из условия устойчивости этой системы следует, что

$$||z|| \le C_1||\psi_h|| + C_2||\eta_h|| \le Ah^k = O(h^k)$$

А это и означает, что

$$||y - u|| = O(h^k)$$

Перейдем к решению конкретных задач.

15.5. Задача построения сеточной аппроксимации

Начнем с одномерной задачи с двумя переменными $x \in [x_0, x_N]$ и $t \in [0, T]$. Введем сетку $\Omega_h = \omega_h + \omega_\tau$. Сетки по отдельным переменным строятся путем разбиения отрезка на которых определены функции

$$\omega_h = \left\{0 = x_0 x_n = x_0 + nh, \quad h = \frac{x_N - x_0}{N}\right\},\ \omega_\tau = \left\{0 = t_0 t_m = t_0 + mh, \quad h = \frac{T - t_0}{M}\right\}$$

Замечание. Индексы h в величинах ω_h и ω_h имеют различные значения. h в омега большом является общей характеристикой сетки построенной в области $\overline{D} = [0, l] \times [0, T]$ и зависит от двух переменных h и τ.

На рисунке 15.3 изображена сеточная область, на которой мы решаем задачу разностную

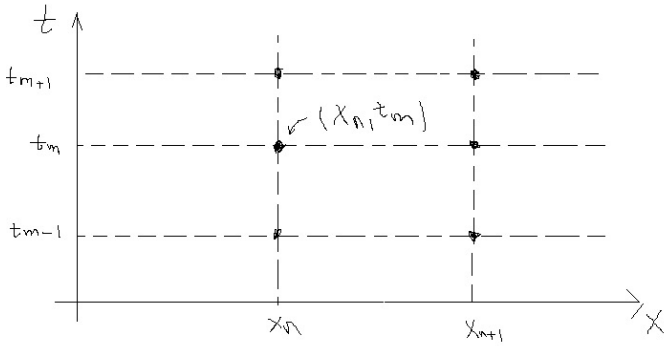

Рис. 15.3 – разбиение двумерной области

(54,55). Часть узлов удовлетворяют первому уравнению для разностной задачи, это внутренние точки, часть удовлетворяют граничным условиям и начальным условиям при $t = t_0$. Причем заметим, что граничные условия, в общем случае, меняются при переходе на другой слой(движение от t_k к t_{k+1}). Сеточную функцию от аргумента x_n, t_m будем обозначать $y = y_n^m = y(x_n, t_m)$.

Построим сеточные аппроксимации для дифференциальных операторов

$$u(x), u'(x), u''(x), \dots$$

и для функций двух переменных

$$u(x,t), u_x(x,t), u_t(x,t)$$

Для функции непрерывных переменных дифференциальный оператор выглядел следующим образом

$$Lu = \frac{d}{dx} i \lim \frac{u(x + \triangle x)}{\triangle x} = u'(x)$$

Построим сеточный аналоги простейших дифференциального оператора первого порядка

$$l_h = \frac{y_{i+1} - y_i}{h}$$

Посмотрим насколько хорошо написанное выражение аппроксимирует функцию $u'(x)$. Для этого найдем невязку

$$L_h u - \underbrace{(u'(x))_h}_{(Lu)_h} = \frac{u(x_i + h) - u(x_i)}{h} - u'(x_i) =$$

$$= \frac{u(x_i) + hu'(x_i) + \frac{h^2}{2}u''(x_i) + o(h^2) - u(x_i)}{h} - u'(x_i) = \frac{h^2}{2}u''(x_i) + o(h^2) = o(h)$$

где $(u'(x))_h$ - производная спроецированная на сетку. Итак, мы видим, что сеточный оператор с первым порядком по h аппроксимирует дифференциальный оператор $\frac{d}{dx}$.

Выпишем формулу Тейлора, опираясь на которую можно было бы строить разложение для различного типа производных

$$u(x+h) = u(x) + hu'(x) + \frac{h^2}{2}u''(x) + \frac{h^3}{3!}u'''(x) + \frac{h^4}{4!}u^{(4)}(x) + o(h^4)$$

$$u(x+2h) = u(x) + 2hu'(x) + \frac{(2h)^2}{2}u''(x) + \frac{(2h)^3}{3!}u'''(x) + \frac{(2h)^4}{4!}u^{(4)}(x) + o(h^4)$$

Аналогично можно написать разложение для $u - h$ и $u - 2h$. Попробуем написать отсюда формулу опираясь на разложение $u \pm h$

$$u(x+h) - u(x-h) = 2hu'(x) + O(h^3)$$

Тем самым $u'(x)$ можно представить

$$u'(x) = \frac{u(x+h) - u(x-h)}{2h} + O(h^2)$$

Поэтому

$$\begin{aligned} L_{\overset{0}{x}} y &= \frac{y_{i+1} - y_{y-1}}{2h}, && \text{центральная производная} \\ L_x y &= \frac{y_{i+1} - y_i}{h}, && \text{производная вперед} \\ L_{\overline{x}} y &= \frac{y_i - y_{i-1}}{h}, && \text{производная назад} \end{aligned}$$

Производную второго порядка $L = \frac{d^2}{dx^2} = \frac{d}{dx}\left(\frac{d}{dx}\right)$ найдем последовательным дифференцированием. Вычислим производную вперед от производной назад

$$L_{\overline{x}x} y = \frac{L_{\overline{x}} y_{i+1} - L_{\overline{x}y_i}}{h} = \frac{\frac{y_{i+1}-y_i}{h} - \frac{y_i - y_{i-1}}{h}}{h} = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2}$$

Посмотрим, что из себя представляет невязка найденного оператора

$$L_{\overline{x}xu-(Lu)_h} = \frac{u(x_i+h) - 2u(x_i) + u(x_i-h)}{h^2} - u''(x_i) = \frac{h^2 u''(x_i) + O(h^4)}{h^2} - u''(x_i)$$

Аналогичным образом строятся производные сеточных операторов по t.

16. Лекция 16. Разностные методы решения задач математической физики. Часть 2

Мы рассмотрели теоретическую часть, описывающую разностные методы. Теперь проиллюстрируем применение этих методов к задачам.

16.1. Одномерное уравнение теплопроводности

Рассмотрим задачу о распространении тепла на отрезке $[0, L]$

$$u_t = a^2 + u_{xx} + f(x,t), \quad 0 < x < l, \quad 0 < t \leq T$$

Дополнительные условия

$$u(x,0) = g_1(x)$$

Краевые условия

$$u(0,t) = g_2(t); \quad u(l,t) = g_3(t)$$

Введем равномерную разностную сетку

$$\Omega = \omega_h \times \omega_\tau, \quad x_n = nh, \quad \tau_m = m\tau, \quad h = \frac{2}{N}, \quad \tau = \frac{T}{M}$$

Узлы сеток нумеруются от 0 до N и M соответственно. Сеточная функция на текущем временном слое - $y(x_n, \tau_m) = y$, на следующем - $\hat{y} = y(x_n, \tau_{m+1})$.

Для аппроксимации первой производной нам нужны две точки, для второй производной три точки. Подходящий шаблон для аппроксимации дифференциальных операторов может быть следующего вида (рис.16.1a)). Этот шаблон называют явной схемой или предельным случаем схемы с весами.

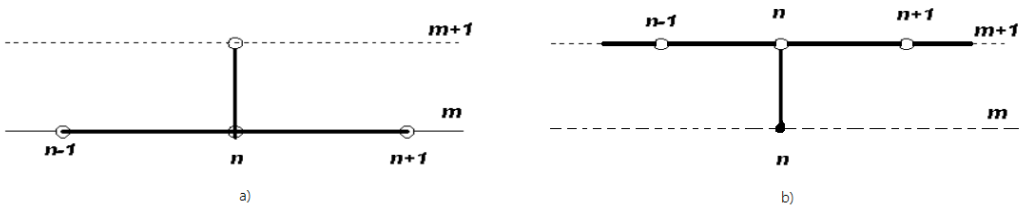

Рис. 16.1 – a) - явная схема, b) - неявная схема

Точно также можно было выбрать шаблон в котором аппроксимация пространственной производной осуществляется на $m+1$ слое, а временной на m и $m+1$(рис. 16.1b)). Этот шаблон называют неявной схемой или предельным случаем схемы с весами. Объединим эти два подхода в шеститочечный шаблон(схема с весами в общем виде) (рис.16.2), впоследствии мы покажем, что он является наиболее удобным для аппроксимации. Для аппроксимации пространственной производной мы будем использовать сеточную функцию с весом $0 \leq \sigma \leq 1$

$$y^{(\sigma)} = \sigma\hat{y} + (1-\sigma)y$$

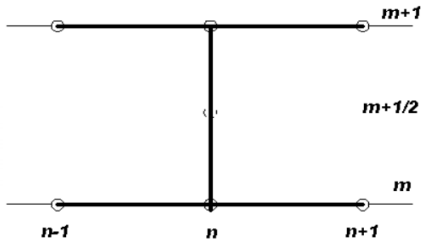

Рис. 16.2 – схема с весами в общем виде

Аппроксимация производной по t

$$\frac{\hat{y}-y}{\tau} = a^2\Lambda_1 y^{(\sigma)} + \varphi_n^m \tag{57}$$

Где $\Lambda_1 = \Lambda_{\overline{x}x}$ сеточный оператор, в котором дифференцирование ведется по первой переменной. Далее запишем разностные условия на границе области, куда включаются начальные при $t = 0$ и краевые условия

$$\begin{aligned} y_n^0 = \chi_1^n = g_1(x_n); \quad & \eta_1 \equiv 0 \\ y_0^m = \chi_2^m = g_2(t_m); \quad & \eta_2 \equiv 0 \\ y_N^m = \chi_3 = g_3(t_m); \quad & \eta_2 \equiv 0 \end{aligned}$$

Изучим невязку уравнения (57). Добавим в наш шаблон фиктивный промежуточный слой $m + \frac{1}{2}$, значение которое на нем принимает t обозначим $\overline{t} = t_m + \frac{\tau}{2}$ и значение u обозначим как $\overline{u} = u\left(x_m, t_m + \frac{1}{2}\right)$. Заметим, что в дискретный набор точек на котором определена функция y не входит точка $(x_n, t_{m+\frac{1}{2}})$, тем не менее решать разностную задачу мы в ней можем.

$$u^{(\sigma)} = \sigma u(x_n, t_m + \tau) + (1-\sigma)u(x_n, t_m) = \sigma u\left(x_n, \overline{t} + \frac{\tau}{2}\right) + (1-\sigma)u\left(x_n, \overline{t} - \frac{\tau}{2}\right) =$$

$$= \sigma\left[u(x_n, \overline{t}) + \overline{u}_t\frac{\tau}{2} + \underbrace{\frac{1}{2!}\overline{u}_{tt}\left(\frac{\tau}{2}\right)^2 + O(\tau^3)}_{O(\tau^2)}\right] + (1-\sigma)\left[u(x_n, \overline{t}) - \overline{u}_t\frac{\tau}{2} + O(\tau^2)\right] =$$

$$= \underbrace{u(x_n, \overline{t})}_{\overline{u}} + \tau\overline{u}_t\left(\frac{\sigma}{2} + \frac{\sigma}{2} - \frac{1}{2}\right) + O(\tau^2)$$

Теперь вернемся к нашей разностной задаче. Подставим в левую часть уравнения (57) непрерывную функцию u. И учтем, что нашу производную вперед по t относительно промежуточного временного можно рассматривать как центральную производную

$$\frac{u(x_n, t_m + \tau) - u(x_n, t_m)}{2 \cdot \frac{\tau}{2}} = \frac{u(x_n, \overline{t} + \frac{\tau}{2}) - u(x_n, \overline{t} - \frac{\tau}{2})}{2 \cdot \frac{\tau}{2}} = u_t(x_n, \overline{t}) + O(\tau^2)$$

Теперь учитывая полученное выражение, подставим функцию u в правую часть уравнения (57) и получим

$$\overline{u}_t+O(\tau^2)=a^2\Lambda_1[\overline{u}+\tau\left(\sigma-\frac{1}{2}\right)\overline{u}_t]+o(\tau^2)]+\varphi_n^m=a^2\overline{u}_{xx}+O(h^2)+\tau\left(\sigma-\frac{1}{2}\right)\overline{u}_{txx}+O(\tau^2)+\varphi_n^m$$

где Λ сеточный аналог оператора Лапласа.

Определим неизвестные величины

1)

$$\varphi_n^m=f(x_n,\bar{t})=f\left(x_n,t_m+\frac{\tau}{2}\right)$$

2)

$$\varphi_n=\begin{cases}O(\tau+h^2), & \sigma\neq\frac{1}{2}\\ O(\tau^2+h^2), & \sigma=\frac{1}{2}\end{cases}$$

Из этого следует, что схема с $\sigma=\frac{1}{2}$ наиболее выгодна с точки зрения аппроксимации. Для явной схемы (рис. 16.1a))$\sigma=0$, для неявной схемы (рис. 16.1b)) $\sigma=1$.

Итак, мы построили разностную схему с указанными порядками аппроксимации, и в случае устойчивости задачи по входным данным, мы можем утверждать, что у схемы будет обладать сходимостью с точностью не ниже порядка аппроксимации.

16.2. Устойчивость разностной схемы

Введем аналоги норм для сеточных функций

$$||y||_c=\max_{n,m}|y_n^m|$$

c-норма на текущем временном слое

$$||y^m||_c=\max_{h}|y_n^m|$$

Аналог среднеквадратичной нормы

$$||y^m||_{l2}=h\sum_{n=0}^{N-1}y_n^2$$

Напомним достаточное условие устойчивости схемы

$$||y^{m-1}||\leq(1-C_1\tau)||y_m||+\tau C_2||\varphi||$$

Относительно легко изучить устойчивость следующих схем.

1) Чисто неявная схема $\sigma=1$

$$y_n^{m+1}-y_n^m=\underbrace{\frac{\tau a^2}{h^2}}_{=\gamma}(y_{n-1}^{m+1}-2y_n^{m+1}+y_{n+1}^{m+1})+\tau\varphi_n^m$$

Предположим, что на слое $m+1$ сеточная функция y_n^{m+1} достигает своего максимального значения в узле k_0, тогда

$$y_n^{m+1} \leq y_{k_0}^{m+1} = y_{k_0}^m - 2(2y_{k_0}^{m+1} - y_{k_0-1}^{m+1} - y_{k_0+1}^{m+1}) + \tau\varphi_{k_0}^m \leq \max_n y_n^m + \tau \max_{n,m} \varphi_n^m$$

Пусть l_0 отвечает минимальному по n значению y_{m+1} на $m+1$-ом временном слое, тогда

$$y_n^{m+1} \geq y_{l_0}^{m+1} = y_{l_0}^m - (2y_{l_0}^{m+1} - y_{l_0-1}^{m+1} - y_{l_0+1}^{m+1})\tau\varphi \geq \min_n y_n^{m+1} + \tau \min_{n,m} \varphi_n^m$$

Таким образом

$$\max_n |y_n^m + 1| = ||y^{m+1}|| \leq ||y^m|| + \tau||\varphi||$$

Мы получили что неявная схема с $\sigma = 1$ безусловно устойчива.

1) Чисто явная схема $\sigma = 0$. При $\sigma = 0$ получаем

$$\frac{\hat{y} - y}{\tau} = a^2\Lambda_1 y + \varphi_n^m$$

На новом временном слое находится одно значение y

$$y_n^{m+1} = \gamma(y_{n-1}^m - 2y_n^m + y_{n+1}^m) + y_n^m + \tau\varphi_n^m \Rightarrow$$
$$y_n^{m+1} = (1-2\gamma)y_n^m + \gamma y_{n-1}^m + \gamma y_{n+1}^m + \tau\varphi_n^m$$

Рассмотрим случай когда $(1-2\gamma) \geq 0 \Rightarrow \gamma \leq \frac{1}{2}$

$$||y_n^{m+1}|| \leq (1-2\gamma)|y_n^m| + \gamma|y_{n-1}^m| + \gamma|y_{n+1}^m| + \tau|\varphi_n^m|$$

$$||y^{m+1}|| \leq 1 \cdot ||y^m||_c + \tau||\varphi||$$

Поэтому $\gamma \leq \frac{1}{2}$ условие устойчивости явной схемы

Соответствующее ограничение на шаги сетки $\gamma = \frac{\tau a^2}{h^2} \leq \frac{1}{2}$. Из этого следует, что $\tau \sim h^2 \sim \frac{1}{N^2}$. Если нам нужно добраться по шагу t до порядка величины $O(1)$, нам необходимо сделать N^2 шагов по времени t. Посмотрим, что будет происходить со схемой при $\gamma > \frac{1}{2}$.

Схема для погрешности

$$\delta y_n^{m+1} = (1-2\gamma)\delta y_n^m + \gamma\delta y_{n-1}^m + \gamma\delta y_{n+1}^m + \tau\varphi_n^m$$

Пусть погрешность быстро осциллирующая на сетке функция и имеет вид

$$\delta y_n^m = (-1)^n\varepsilon$$

Тогда на следующем слое $m+1$

$$\delta y_n^{m+1} = (1-2\gamma)\varepsilon(-1)^n + \gamma\varepsilon(-1)^{n-1} + \gamma\varepsilon(-1)^{n+1} =$$
$$= \varepsilon(-1)^n(1 - 2\gamma - \gamma - \gamma) = (-1)^{n+1}(4\gamma - 1)\varepsilon$$

Это геометрическая прогрессия со знаменателем больше единицы

$$||\delta y^{m+k}|| = (4y-1)^k \varepsilon \to \infty, \quad k \to \infty$$

таким образом, несмотря на то, что норма нашей погрешности в C(это величина ε) мала, тем не менее вычисления на схеме неустойчивы и с ростом индекса N стремиться к бесконечности.

Схема с весами в общем случае

Рассматриваются гармоники y степени роста m разностной задачи

$$(y_k)_n^m = (\rho_k)^m e^{ikx_n}$$

Гармоника на следующем временном слое будет отличаться от предыдущего временного слоя на ρ

$$\hat{y} = \rho_k y$$

Достаточное условие устойчивости

$$|\rho_k| \leq (1 - C\tau) \quad \forall k$$

Для практических задач чаще пользуются более жестким условием

$$|\rho_k| \leq 1$$

его мы и будем применять.

Используем метод гармоник применительно к нашей схеме с однородными уравнениями

$$\hat{y} - y = a^2 \Lambda_1(\sigma\hat{y} + (1-\sigma)y))$$

$$(\rho-1)y = \rho_k^m \frac{\tau a^2}{h^2}[\sigma\rho_k(e^{ik(x_n-h)} - 2e^{ikx_n} + e^{ik(x_n+h)}) + (1-\sigma)(e^{ik(x_n-h)} - 2e^{ikx_n} + e^{ik(x_n+h)})] =$$

$$= \gamma^2 y\left[\sigma\rho_k(2\cos(kh) - 2) + (1-\sigma)\left(-4\sin^2\left(\frac{kh}{2}\right)\right)\right] \Rightarrow$$

$$\rho_k - 1 = \gamma\sigma\rho_k\left(-4\sin^2\left(\frac{kh}{2}\right)\right) + \gamma(1-\sigma)\left(-4\sin^2\left(\frac{kh}{2}\right)\right) =$$

$$\rho_k\left(1 + \gamma\sigma 4\sin^2\left(\frac{kh}{2}\right)\right) = 1 + \gamma\sigma 4\sin^2\left(\frac{kh}{2}\right) - \gamma 4\sin^2\left(\frac{kh}{2}\right)$$

Получаем соотношение для ρ_k

$$-1 \leq \rho_k = \frac{1 + \gamma\sigma 4\sin^2\left(\frac{kh}{2}\right) - \gamma 4\sin^2\left(\frac{kh}{2}\right)}{1 + \gamma\sigma 4\sin^2\left(\frac{kh}{2}\right)} \leq 1$$

Очевидно выполнение неравенства $\rho_k \leq 1$, разберемся с оставшимся неравенством

$$0 \leq 2 + 2\gamma\sigma 4\sin^2\left(\frac{kh}{2}\right) - \gamma 4\sin^2\left(\frac{kh}{2}\right)$$

$$2\gamma\sigma 4\sin^2\left(\frac{kh}{2}\right) \geq -2 + \gamma 4\sin^2\left(\frac{kh}{2}\right)$$

$$\sigma \geq \frac{1}{2} - \frac{1}{4\gamma\sin^2\left(\frac{kh}{2}\right)} \Rightarrow$$

$$\sigma \geq \frac{1}{2} - \frac{1}{4\gamma}$$

Полученное неравенство называется условием Куранта(для среднеквадратичной нормы). Тем самым мы исследовали все σ от нуля до единицы.

Для устойчивости в равномерной нормы условие устойчивости

$$\sigma \geq 1 - \frac{1}{2\gamma}$$

Запишем как выглядит решение для простейших схем

1) Явная схема $\sigma = 0$

$$y_n^{m+1} = y_n^m + \gamma(y_{n-1}^m - 2y_n^m + y_{n+1}^m) + \varphi_n^m$$

Характер вычислений имеет явный характер по индексу n

2) Неявная схема $\sigma = 1$

$$y_n^m + \tau\varphi_n^m = -\gamma y_{n-1}^{m+1} + (1+2\gamma)y_n^{m+1} - \gamma y_{n+1}^{m+1}$$

16.3. Разностная схема крест

Рассмотрим задачу для уравнений колебаний с краевыми условиями Дирихле

$$\begin{cases} u_{tt} = a^2 u_{xx} + f(x,t), \quad 0u(x,0) = g_1(x), \qquad 0u_t(x,0) = g_2(x) \\ u(0,t) = g_3(t) \\ u(l,t) = g_u(t) \end{cases}$$

Нам требуется аппроксимация второй производной поэтому необходимы три временных слоя. Предложим простейший шаблон рис.(16.3).

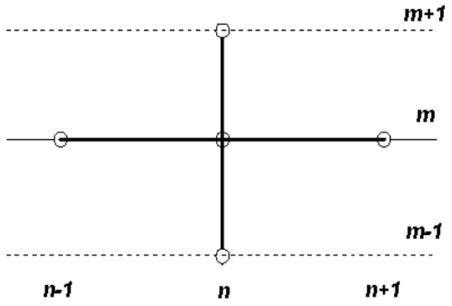

Рис. 16.3 – схема крест

Построим сеточную аппроксимацию

$$\frac{\hat{y} - 2y + \overset{\vee}{y}}{\tau} = \frac{y_n^{m+1} - 2y_n^m + y_n^{m-1}}{h^2} + \varphi_n^m = a^2 \frac{y_{n-1}^m - 2y_n^m + y_{n+1}^m}{h^2} + \varphi_n^m$$

При подстановке непрерывного решения в построенную разностную схему

$$u_{tt}(x_n, t_m) + O(\tau^2) = a^2 u_{xx}(x_n, t_m) + O(h^2) + \varphi_n^m$$

Для аппроксимации нашего уравнения нам нужно отождествить $\varphi_n^m = f(x_n, t_m)$, тогда невязка $\Psi_h = O(\tau^2 + h^2)$. Теперь получим невязку связанную с аппроксимацией входных данных

$$\begin{aligned} y_n^0 = \chi_1^n = g_1(x_n) \quad &\Rightarrow \quad \eta_1 = 0 \\ y_0^m = \chi_3^m = g_3(t_m) \quad &\Rightarrow \quad \eta_3 = 0 \\ y_N^m = \chi_4^m = g_4(t_m) \quad &\Rightarrow \quad \eta_4 = 0 \end{aligned}$$

Напишем аппроксимацию для u_t

$$\frac{y_n^1 - y_n^0}{\tau} = \chi$$

17. Лекция 17. Разностные методы решения задач математической физики. Часть 3

Изучим устойчивость схемы крест.

Напомним, что мы рассматриваем однородную задачу

$$\hat{y} - 2y + \overset{\vee}{y} = \frac{\tau^2 a^2}{h^2}(y^m_{n-1} - 2y^m_n + y^m_{n+1})$$

Напомним, соотношения между слоями для гармоник $y_k = \rho^m_k e^{ikx_n}$

$$\hat{y} = \rho y, \quad \overset{\vee}{y} = \frac{1}{\rho} y$$

Тогда перепишем нашу задачу в виде

$$y\left(\rho - 2 + \frac{1}{\rho}\right) = \underbrace{\frac{\tau^2 a^2}{h^2}}_{\gamma^2} y\left(- 4\sin^2\left(\frac{kh}{2}\right)\right) \Rightarrow$$

$$\rho^2 - 2\rho\left(1 - \frac{\gamma^2}{2} 4\sin^2\left(\frac{kh}{2}\right)\right) + 1 = 0$$

Мы получили квадратное уравнение относительно ρ. Из теоремы Виета видно, что $\rho_1 \cdot \rho_2 = 1$. Из условия устойчивости нам нужно чтобы $||\rho|| < 0$. Отсюда мы делаем вывод, что модули корней должны совпадать, а значит ρ_1, ρ_2 - комплексно сопряженные корни. Для невырожденных случаев, когда корень имеет кратность два, дискриминант должен быть меньше нуля

$$\left(1 - 2\gamma^2 \sin^2\left(\frac{kh}{2}\right)\right)^2 - 1 < 0 \Leftrightarrow$$

$$\left|1 - 2\gamma^2 \sin^2\left(\frac{kh}{2}\right)\right| < 1$$

из этого получаем условие Куранта для $\gamma(\tau$ и $h)$

$$\gamma = \frac{\tau a}{h} < 1 \tag{58}$$

1) Схема "крест" устойчива в среднем по начальным данным при дополнительном условии $\frac{\tau a}{h} < 1$.

2) При условии (58) схема "крест" устойчива по правой части;

3) При условии (58) схема "крест" устойчива по начальным данным и правой части в равномерной сеточной норме (в C).

17.1. Многомерные разностные схемы для уравнения теплопроводности

Рассмотрим задачу о распределении тепла в прямоугольной области:

$$\begin{cases} u_t = a^2(u_{x_1x_1} + u_{x_2x_2} + f(x_1, x_2, t)) & 0 < x_1 < l_1 \\ u|_\Gamma = g(x_1, x_2, t), \ \ (x_1, x_2) \in \Gamma & 0 < x_2 < l_2 \\ u(x_1, x_2, 0) = g(x_1, x_2) & 0 < t \le T \end{cases}$$

Равномерные сетки (рис.17.2) строятся по полной аналогии с одномерным случаем. Сеточные функции записываются как

$$y(x_{1n}, x_{2k}, y_m) = y^m_{n,k} \equiv y$$

Построим разностную схему

$$\frac{\hat{y} - y}{\tau} = a^2(\Lambda_1 + \Lambda_2)y^{(\sigma)} + \varphi^m_{n,k} \tag{59}$$

Напомним, что $\Lambda_1 = \Lambda_{\overline{x}_1,x_1}$ и $\Lambda_2 = \Lambda_{\overline{x}_2,x_2}$ - сеточные дифференциальные операторы Лапласа, продолжим

$$y^0_{n,k} = \chi_{n,k} = g(x_{1n}, x_{2k}), \quad \eta = 0$$

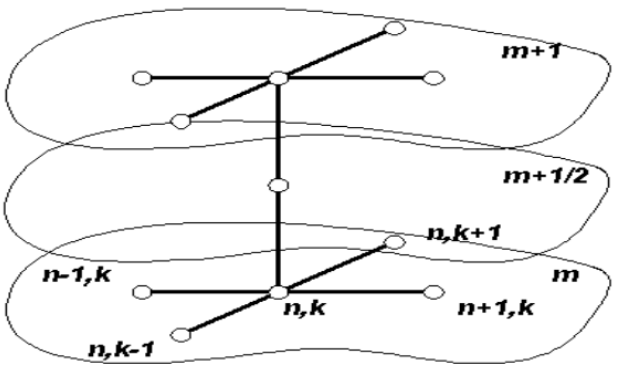

Рис. 17.1 – разностная многомерного случая

Невязка равна нулю, мы точно аппроксимируем значения непрерывной функции в узлах сетки. Просьба показать что $\eta_\Gamma = 0$. Основной проблемой для нас остается аппроксимация основного уравнения (59) для того, чтобы определить порядок аппроксимации нашей разностной схемы. Для этого нам нужно записать для непрерывной функции представление в виде функции с весом σ, напомним что эта функция вычисляется как

$$u^{(\sigma)} = \delta u\left(x_{1n}, x_{2k}, t_m + \frac{\tau}{2} + \frac{\tau}{2}\right) + (1-\sigma)u\left(x_{1n}, x_{2k}, \overline{t} - \frac{\tau}{2}\right) =$$

$$= \overline{u}\left(x_{1n}, x_{2k}, t_m - \frac{\tau}{2}\right) + \tau\left(\sigma - \frac{1}{2}\right)\overline{u}_t + O(\tau^2)$$

У нас снова выделилась схема повышенного порядка аппроксимации с $\sigma = \frac{1}{2}$. Рассмотрим более подробно как действуют сеточный оператор $\Lambda_1 + \Lambda_2$ на функцию с весом σ

$$(\Lambda_1 + \Lambda_2)y^{(\sigma)} = \sigma\left(\frac{y^{m+1}_{n-1,k} - 2y^{m+1}_{n,k} + y^{m+1}_{n+1,k}}{h_1^2} + \frac{y^{m+1}_{n,k-1} - 2y^{m+1}_{n,k} + y^{m+1}_{n,k+1}}{h_2^2}\right) +$$

$$+(1-\sigma)\left(\frac{y^m_{n-1,k}-2y^m_{n,k}+y^m_{n+1,k}}{h_1^2}+\frac{y^m_{n,k-1}-2y^m_{n,k}+y^m_{n,k+1}}{h_2^2}\right)$$

Итак, у нас есть все, чтобы записать

$$\frac{\hat{u}-u}{2\frac{\tau}{2}}=\overline{u}_t+O(\tau^2)=a^2(\overline{u}_{x_1,x_1}+\overline{u}_{x_2,x_2})+O(h_1^2+h_2^2)+\tau\left(\sigma-\frac{1}{2}\right)(\overline{u}_{tx_1,x_1}+\overline{u}_{tx_2,x_2})+\varphi^m_{n,k}$$

$$\varphi^m_{n,k}=f(x_{1n},x_{2k},t_m+\frac{\tau}{2})$$

слагаемые $\varphi^m_{n,k}$ аннулируются, так как уравнение непрерывно, остается только невязка, которая равна

$$\Psi_n=\begin{cases}O(\tau+h_1^2+h_2^2), & \sigma\neq\frac{1}{2}\\ O(\tau^2+h^2+h_2^2), & \sigma=\frac{1}{2}\end{cases}$$

Чтобы установить устойчивость снова будем пользоваться методом гармоник, запишем как они будут выглядеть

$$(y_{p,q})^m_{n,k}=\rho_{p,q}e^{ipx_{1n}+ipx_{2k}}$$

Как правило, мы не будем перегружать нашу запись индексами p,q, предполагая что при изучении устойчивости в нашу схему подставляются гармоники. Нам нужны только два временных слоя поэтому запишем связь между гармониками на этих уровнях

$$\hat{y}=\rho_{p,q}\equiv\rho y$$

Тогда, аналогично одномерному случаю

$$\rho-1=\tau a^2\left[\rho\sigma\left(\frac{-4\sin^2\left(\frac{\rho h_1}{2}\right)}{h_1^2}+\frac{-4\sin^2\left(\frac{\rho h_2}{2}\right)}{h_2^2}\right)+\right.$$
$$\left.+(1-\sigma)\rho\sigma\left(\frac{-4\sin^2\left(\frac{\rho h_1}{2}\right)}{h_1^2}+\frac{-4\sin^2\left(\frac{\rho h_2}{2}\right)}{h_2^2}\right)\right.$$

Обозначим

$$A=\tau a^2\frac{-4\sin^2\left(\frac{\rho h_1}{2}\right)}{h_1^2}$$
$$B=\tau a^2\frac{-4\sin^2\left(\frac{\rho h_2}{2}\right)}{h_2^2}$$

$$\rho - 1 = \sigma\rho(-A - B) + (1 - \sigma)(-A - B)$$

Необходимо решить уравнение

$$\rho = \frac{1 + \sigma(A + B) - (A + B)}{1 + \sigma\rho(A + B)}$$

И неравенство

$$|\rho| \leq 1$$

Постарайтесь, в качестве самостоятельного упражнения, отсюда получить ограничения на σ, τ, h.

$$\sigma \geq \frac{1}{2} - \frac{1}{4(\gamma_1 + \gamma_2)} \quad \Rightarrow \quad \gamma_1, \gamma_2 \leq \frac{1}{4}$$

Мы установили устойчивость и аппроксимацию, осталось разобраться с задачами вычисления.

Существенный недостаток схемы (59) в многомерном случае связан с тем, что как чисто явная схема $\sigma = 0$, так и неявная $\sigma \neq 0$ схемы приводят к неэффективным численным алгоритмам для построения решения на слое T. Если из соображений аппроксимации $h_1 \sim h_2; N \sim K$, то оценка числа арифметических действий для явной схемы для построения решения на последнем слое есть $O(N4)$. Действительно, для перехода на следующий временной слой решается явная система уравнений с числом неизвестных$O(NK) \sim N^2$. При этом требования устойчивости схемы ограничивают временной шаг $\tau \sim \left(\frac{1}{h^2}\right)^{-1} \sim h^2 \sim N^{-2}$. Что и приводит к общей оценке числа арифметических действий $O(N4)$.

Для неявный схемы положение еще хуже. Ограничиваясь абсолютно устойчивым вариантом схем при $\sigma \geq \frac{1}{2}$ на каждом временном слое приходится решать СЛАУ с уравнений при ширине ленты порядка $O(2N)$. Метод исключения Гаусса требует $O(N^6)$ с учетом ленточной структуры матрицы $O(N^4)$ действий. Требование аппроксимации дает $O(N)$ шагов по времени. Итого $O(N^5)$ действий. Неявная схема менее выгодна в этом случае.

Поэтому предпочтение отдают абсолютно устойчивым $\tau \sim h$, экономичным разностным схемам, в которых при переходе на очередной временной слой совершается всего $O(N^2)$ действий.

17.2. Продольно-поперечная разностная схема

Введем промежуточный по t слой $(m + \frac{1}{2})$ и рассмотрим разностную схему

$$\begin{aligned} \frac{\overline{y} - y}{\frac{\tau}{2}} &= a^2\Lambda_1\overline{y} + a^2\Lambda_2 y + \overline{f} \\ \frac{\hat{y} - y}{\frac{\tau}{2}} &= a^2\Lambda_1\overline{y} + a^2\Lambda_2\hat{y} + \overline{f} \end{aligned} \tag{60}$$

Обсудим построение решения уравнения (60) на $m + 1$-ом слое:

1) Уравнение (60.1) позволяет найти $\overline{y}_{n,k}$ по неявной схеме относительно x_1 и по явнои схеме относительно x_2 $\Rightarrow$ решается система с 3-х диагональная матрицей

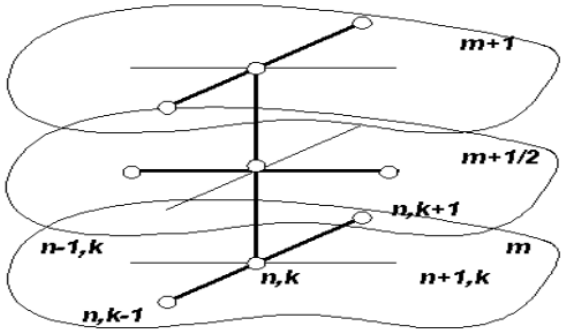

Рис. 17.2 – процесс диффузии

относительно переменной x_1 эффективным методом прогонки по x_1 при каждом k (k - раз прогонка с $O(N) \Rightarrow O(NK)$действий).

2) Уравнение (60.2) позволяет найти $\hat{y}_{n,k}$ по неявной схеме относительно x_2 и по явной схеме относительно $x_1 \Rightarrow$ прогонка по x_2 при каждом $n \Rightarrow O(NK)$ действии $\Rightarrow$ итого $O(2NK) \sim O(N^2)$ действии.

3) Диагональные коэффициенты в соответствующих матрицах на каждом шаге преобладают - тем самым решение существует, единственно и вычисления по формулам прогонки устойчивы.

4) Общее число действий при переходе на $(m+1)$-ый временной слой $O(30N^2)$ действии.

17.3. Устойчивость продольно-поперечной неявной схемы

Будем использовать метод гармоник

$$y_{n,k} = exp(ix_{1n}p + ix_{2k}q); \quad \overline{y} = p'_{p,q}y; \quad \hat{y} = p''_{p,q}\overline{y}$$

свои множители роста на каждой полуполосе

$$p'_{p,q} - 1 = \frac{\tau a^2}{2h_1^2}\left(-4\sin^2\frac{ph_1}{2}\right)p'_{p,q} + \frac{\tau a^2}{2h_2^2}\left(-4\sin^2\frac{qh_2}{2}\right)$$

Аналогично получется уравнение для $p''_{p,q}$. Из этих уравнений получаем

$$\begin{cases} p'_{p,q} = \frac{1-\frac{2\tau a^2}{h_2^2}\sin^2\frac{qh_2}{2}}{1+\frac{2\tau a^2}{h_1^2}\sin^2\frac{ph_1}{2}} \\ p''_{p,q} = \frac{1-\frac{2\tau a^2}{h_1^2}\sin^2\frac{ph_1}{2}}{1+\frac{2\tau a^2}{h_2^2}\sin^2\frac{qh_2}{2}} \end{cases} \quad \Rightarrow \quad |p_{p,q}| = |p'_{p,q}p''_{p,q}| \leq 1$$

всегда, то есть $\forall p$ и q. таким образом, схема безусловно (абсолютно) устойчива по начальным данным (и по правой части тоже)

Осталось установить аппроксимацию.

17.4. Аппроксимация продольно-поперечной схемы

Исключим из (60) слой $\overline{y}_{n,k}$. Для этого вычтем уравнения (1)-(2), найдем

$$2\frac{\overline{y}_{n,k}}{\tau/2} - \frac{\hat{y}_{n,k} + y_{n,k}}{\tau/2} = -a^2\Lambda_2(\hat{y} - y),$$

$$\overline{y}_{n,k} = \frac{\hat{y}_{n,k} + y}{2} - \frac{\tau a^2}{4}\Lambda_2(\hat{y} - y) \tag{61}$$

Складывая уравнения (1) и (2), найдем:

$$\frac{\hat{y}_{n,k} - y_{n,k}}{\tau/2} = a^2\Lambda_1(2\overline{y}_{n,k}) + a^2\Lambda_2(\hat{y}_{n,k} + y_{n,k}) + 2\overline{f}$$

Откуда с учетом (61), получим

$$\frac{\hat{y}_{n,k} - y_{n,k}}{\tau} = a^2\Lambda_1\frac{\hat{y}_{n,k} + y_{n,k}}{2} - \frac{\tau a^2}{4}\Lambda_1\Lambda_2(\hat{y} - y) + a^2\Lambda_2\frac{\hat{y}_{n,k} + y_{n,k}}{2} + \overline{f} =$$

$$= a^2(\Lambda_1 + \Lambda_2)\frac{\hat{y}_{n,k} + y_{n,k}}{2} - \underbrace{\frac{\tau a^2}{4}\Lambda_1\Lambda_2(\underbrace{\hat{y} - y}_{O(\tau)})}_{O(\tau^2)} + \overline{f}$$

Это почти симметричная схема с $\sigma_1 = \sigma_2 = \frac{1}{2}$, тем самым - схема обладает аппроксимацией при условии $\overline{f} = f(x_{1n}, x_{2n}, t_{m+\frac{1}{2}})$ и порядок аппроксимации

$$\psi = O(\tau^2 + h_1^2 + h_2^2)$$

Схема (60) безусловно устойчива и обладает повышенной аппроксимацией, следовательно, она сходится в указанной прямоугольной области на равномерной сетке и обладает точностью не хуже, чем

$$||y - u|| = O(\tau^2 + h_1^2 + h_2^2)$$

18. Лекция 18. Дополнительная

18.1. Постановка задачи

Сделаем некоторые замечания к нашим последним лекциям на которых мы рассматривали разностные схемы для уравнения теплопроводности для $1D$ и $2D$ областей. В общем виде уравнение теплопроводности записывается как

$$\rho c_p \frac{\partial T}{\partial t} = \frac{\partial}{\partial x}\left(k_z \frac{\partial T}{\partial z}\right) + \frac{\partial}{\partial y}\left(k_y \frac{\partial T}{\partial y}\right) + f(x, y, t)$$

Приблизительный процесс диффузии проиллюстрирован на рис.18.1.

Рис. 18.1 – процесс диффузии

Мы ограничимся простейшим случаем для прямоугольной области без дополнительных источников в области, а соответствующий режим будет определяться краевыми и начальными условиями, уравнение записывается в виде

$$\frac{\partial T}{\partial t} = \alpha\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2}\right) \tag{62}$$

где, $\alpha = \frac{k}{\rho c_p}$.

За этим стоит следующая физическая модель (рис.18.2) Это распределения тепла в некоторой прямоугольной площадки, в каком-то смысле это моделирование нагрева некоторого электронного чипа. На двух границах заданы условия Дирихле с соответствующей температурой, на других двух границах заданы тепловые потоки(условия Неймана), и начальная температура 20 градусов по цельсию. Стандартная сетка которую мы используем представлена на рис.18.3. Роль дискретной функции u играет функция T имеющая смысл температуры. Вторые производные аппроксимированы с помощью трех узлов, шаблон который при этом используется

$$\frac{T_{i,j}^{n+1} - T_{i,j}^{n}}{\triangle t} = \alpha\left(\frac{T_{i+1,j}^{n} - 2T_{i,j}^{n} + T_{i-1,j}^{n}}{\triangle x^2} + \frac{T_{i,j+1}^{n} - 2T_{i,j}^{n} + T_{i,j-1}^{n}}{\triangle y^2}\right) \tag{63}$$

Шаги h_x, h_y и τ обозначены как $\triangle x^2$, $\triangle y^2$ и $\triangle \tau$ соответственно.

18.2. Явная схема

Если рассматривать простейшую явную схему, то вычисления на новом временном слое $n+1$ присутствуют только лишь в одной сеточной переменной, отнесенной

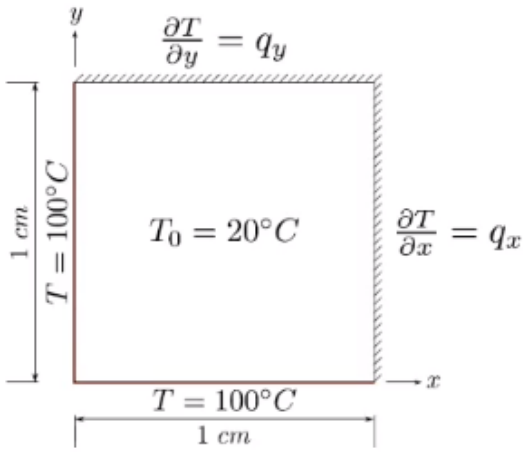

Рис. 18.2 – физическая модель

к центральному узлу. Таким образом, для вычисления нового значения на $n+1$-ом слое мы получаем простое уравнение

$$T_{i,j}^{n+1} = T_{i,j}^{n} + \alpha\left(\frac{\triangle t}{\triangle x^2}(T_{i+1,j}^{n} - 2T_{i,j}^{n} + T_{i-1,j}^{n}) + \frac{\triangle t}{\triangle y^2}(T_{i,j+1}^{n} - 2T_{i,j}^{n} + T_{i,j-1}^{n})\right)$$

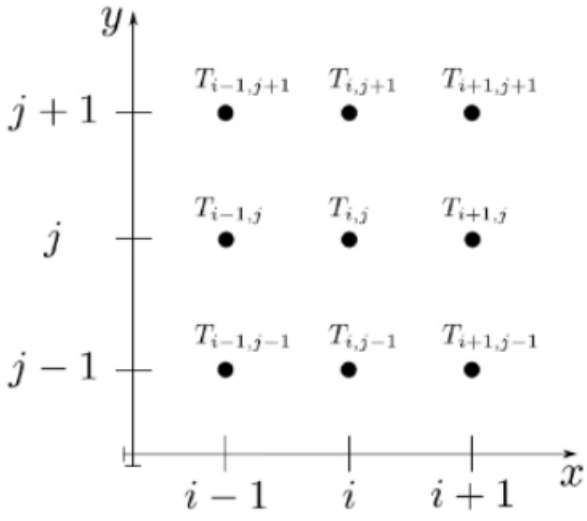

Рис. 18.3 – узлы разностной схемы

Тем самым нам не требуется составление системы уравнений, мы можем для текущего центрального узла с индексом (i, j) явно вычислить значения T на новом

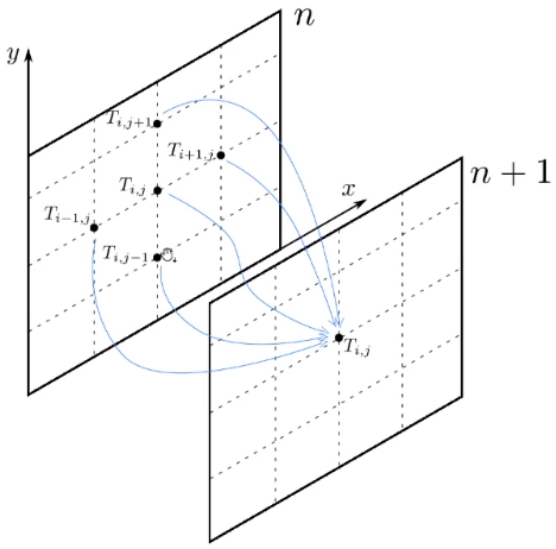

Рис. 18.4 – разностная схема

временном слое, владея прежними значениями распределенными на соответствующем пространственном шаблоне. Схематически это изображено на рис.18.4

Аппроксимация краевых условий

$$T_{i,end} = q_y \triangle y + T_{i,end-1}$$
$$T_{end,i} = q_z \triangle x + T_{end-1,i}$$

Ограничение установленное для общей схем с весами(и в частности для явной схемы)

$$\alpha \frac{\triangle t}{(\triangle x)^2} + \alpha \frac{\triangle t}{(\triangle x)^2} < 2$$

в простейшем случае при $\triangle x = \triangle y = \delta$

$$\alpha \frac{\triangle t}{\delta^2} < \frac{1}{4}$$

Таким образом, мы понимаем, что нам не нужны формирование матрицы, решения основной задачи линейной алгебры, мы можем просто работать с двумерным массивом, обновляя его переходя на следующий шаг.

18.3. Неявная схема

Решаем то же самое уравнение (62). Реорганизуем уравнение (63) для неявной схемы $\sigma = 1$, получим

$$-\frac{\alpha \triangle t}{\triangle x^2}(T^{n+1}_{i-1,j} + T^{n+1}_{i+1,j}) + \left(1 + 2\frac{\alpha \triangle t}{\triangle x^2} + 2\frac{\alpha \triangle t}{\triangle y^2}\right) T^{n+1}_{i,j} - \frac{\alpha \triangle t}{\triangle y^2}(T^{n+1}_{i,j-1} + T^{n+1}_{i,j+1}) = T^n_{ij}$$

Если величины шагов по $\triangle x$ и $\triangle y$ одинаковы и равны δ, то мы приходим к системе уровней следующего вида

$$-T^n_{i-1,j} + 1 - T^{n+1}_{i+1,j} + \left(\frac{\delta^2}{\alpha \triangle t} + 4\right) T^{n+1}_{i,j} - T^{n+1}_{i,j-1} - T^{n+1}_{i,j+1} = \frac{\delta^2}{\alpha \triangle t} T^n_{i,j}$$

С краевыми условиями нужно разобраться отдельно, посмотрев во что превращается это уравнение, когда мы попадаем на границу области.

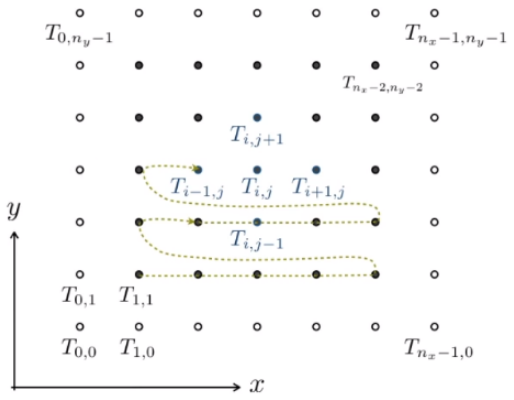

Рис. 18.5 – схема дискретных узлов

Для левой границы

$$-T^{n+1}_{2,j} + \left(\frac{\delta^2}{\alpha \triangle t} + 4\right) T^{n+1}_{1,j} - T^{n+1}_{1,j-1} - T^{n+1}_{1,j-1} - T^{n+1}_{1,j+1} = \frac{\delta^2}{\alpha \triangle t} T^n_{1,j} + T^{n+1}_{0,j}$$

Для правой границы

$$\frac{T^{n+1}_{n_x-1,j} - T^{n+1}_{n_x-2,j}}{\delta} = q_x$$

Переобозначим $T^{n+1}_{n_z-1,j} = \delta q_x + T^{n+1}_{n_z-2,j}$ чтобы найти конечное выражение для последней точки $i = n_x - 2$

$$-T^{n+1}_{n_z-3,j} + \left(\frac{\delta^2}{\alpha \triangle t} + 3\right) T^{n+1}_{n_z-2,j} - T^{n+1}_{n_z-2,j-1} - T^{n+1}_{n_z-2,j+1} = \frac{\delta^2}{\alpha \triangle t} T^n_{n_z-2,j} + \delta q_z$$

Аналогично рассматриваются верхния и нижняя границы.

ФИЗИЧЕСКИЙ
ФАКУЛЬТЕТ
МГУ ИМЕНИ
М.В. ЛОМОНОСОВА

teach-in

ЛЕКЦИИ УЧЕНЫХ МГУ