МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ВВЕДЕНИЕ В ЧИСЛЕННЫЕ МЕТОДЫ

ФИЗИЧЕСКИЙ ФАКУЛЬТЕТ

Московский Государственный университет
им.М.В.Ломоносова

ВВЕДЕНИЕ В ЧИСЛЕННЫЕ МЕТОДЫ

Физический Факультет

Содержание

ГЛАВА I

ВВЕДЕНИЕ

§1. Математическое моделирование. Численные методы и использование ЭВМ в решении прикладных задач

Рассматривая *математический анализ явления* как своего рода *теоретический эксперимент*, из общих и достаточно естественных соображений процесс *математического моделирования* разбивается на несколько этапов:

- **Формулировка математической модели явления.** Математическая модель любого изучаемого явления, по причине его чрезвычайной сложности, должна охватывать важнейшие для рассматриваемой задачи стороны процесса, его существенные характеристики и формализованные связи, подлежащие учёту.

  Как правило, *математическая модель* изучаемого физического явления формулируется в виде *уравнений математической физики*. На этой стадии анализа это существенно нелинейные, многомерные системы уравнений, содержащие большое число неизвестных и параметров.

  Если *математическая модель* выбрана недостаточно тщательно, то какие бы мы не применяли методы для дальнейших расчётов, полученные результаты будут *ненадёжны*, а в отдельных случаях и совершенно *неверны*.

- **Проведение математического исследования** полученной модели и получение соответствующего *решения*.

  На этом этапе моделирования, в зависимости от сложности рассматриваемой модели, применяют различные подходы к её исследованию и различный смысл вкладывается в понятие *решения* задачи. Скажем, доказательство теорем *существования и единственности* в определённом смысле *решает* задачу, однако, являясь зачастую неконструктивным, оно не позволяет нам решить проблему изучения качественного поведения решения и оценки его количественных характеристик.

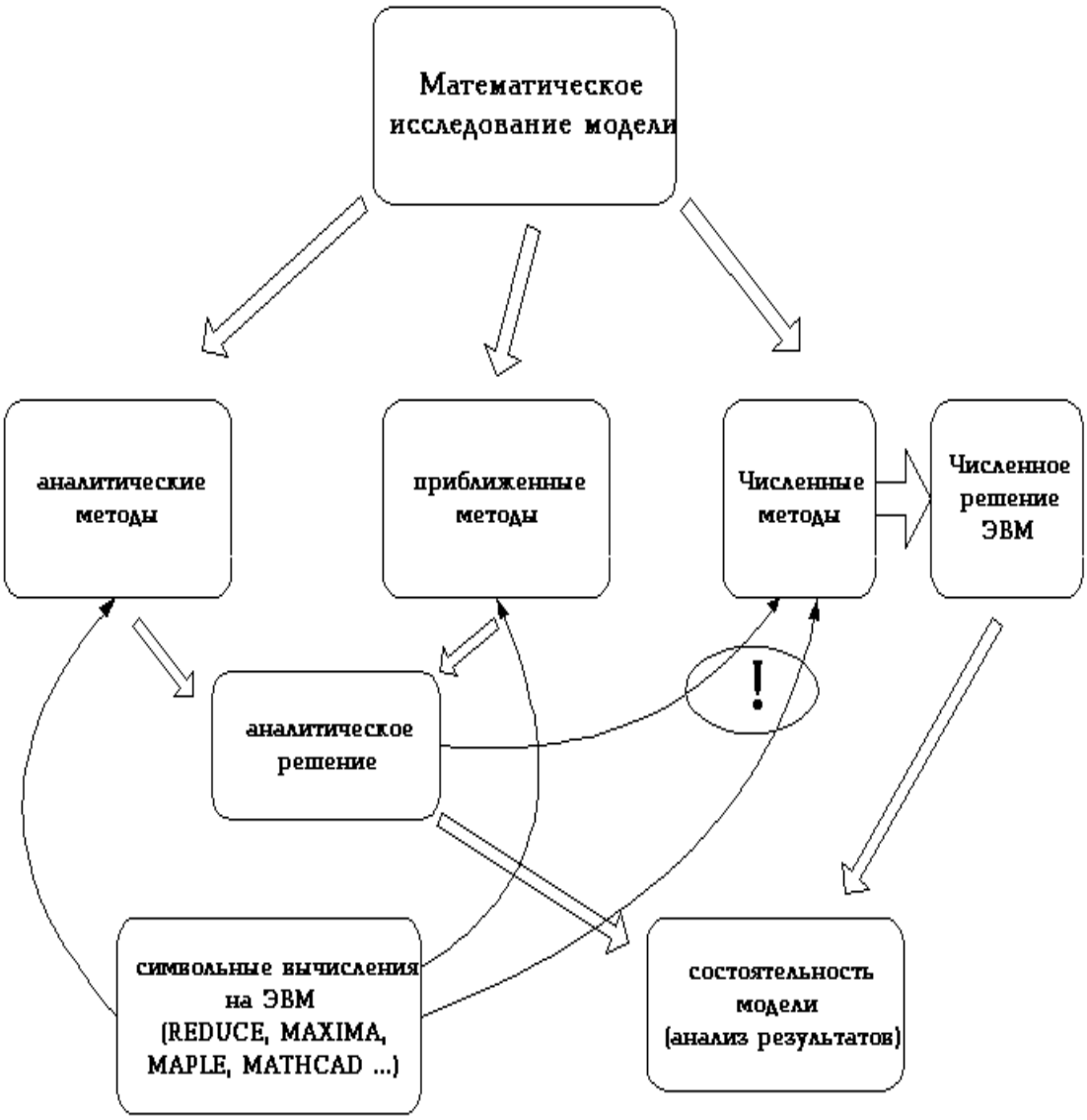
  Для наиболее *грубых* и несложных (в некотором смысле) моделей удаётся получить их *аналитическое решение*. Следует оговориться — использование средств *символьных вычислений* на ЭВМ таких как REDUCE, MAXYMA, MAPLE, "интеллектуальных калькуляторов" MATHEMATICA, MathCAD, MathLab и пр. существенно революционизировало это, традиционное для "бумаги и карандаша", поле деятельности.

  Для более точных и сложных моделей *аналитическое решение* удаётся получить сравнительно редко. При теоретическом анализе задачи в такой ситуации

пользуются обычно *приближенными* математическими методами, например разложением по малому параметру, осреднением, изучение различных асимптотик и другими. Эти приёмы позволяют опять-таки представить приближенное решение в аналитической форме и с его помощью получить удовлетворительные численные результаты.

Наконец для наиболее точных и сложных моделей основными методами решения являются *численные* методы решения с необходимостью требующие проведения большого объёма вычислений на ЭВМ. Эти методы позволяют добиться хорошего *количественного* и даже *качественного* результата в описании модели. Но, правда, у них есть и принципиальные недостатки — как правило, речь идёт о рассмотрении некоторого *частного* решения.

Приведённая схема частично отражает обсуждаемые взаимосвязи этапов математического моделирования.



Как мы видим, каждый из этапов математического исследования модели связан с использованием *численных методов* и получением *численного решения* задачи.

- **Анализ состоятельности предложенной модели,** т. е. осмысление результатов решения, сопоставление полученного решения с имеющимися данными физического эксперимента. На этом этапе решается вопрос о состоятельности математической модели и проведённого исследования. "Хорошее" согласование с "экспериментом" обычно свидетельствует о правильности выбора модели. В противном случае необходимы дополнительные уточнения, изменения и т. п., повторение предыдущих этапов исследования.

Обсуждая предмет лекционного курса, мы акцентировали наше внимание на двух сторонах предмета "Численные методы": *этапе в математическом моделировании* и на *необходимом моменте в процессе исследования, сопряженном с использованием ЭВМ.*

Использование ЭВМ в процессе математического исследования модели требует специфических, численных методов, т.е. такой "интерпретации" математической модели, которая может быть реализована на ЭВМ – назовём её *дискретной* (или *вычислительной*) моделью. Поскольку ЭВМ выполняет только арифметические и логические операции, то для реализации *вычислительной модели* требуется разработка соответствующего *вычислительного алгоритма.* Дальнейшая последовательность действий — это программирование. расчет на ЭВМ, обработка результатов расчета.

В рамках нашего лекционного курса мы остановимся на отдельных проблемах численных методов при анализе сравнительно простых и ставших классическими математических моделей.

Теперь посмотрим на проблему "численных методов" несколько подругому.

§2. Задача "вычисления"

2.1 Задача "вычисления". Анализ постановки

Обычно задачу вычисления величины y по известной величине x записывают, с учётом интересующих нас причинно-следственных связей, в виде

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

где $y \in \mathcal{Y}, \quad x \in \mathcal{X}$ – элементы соответствующих функциональных пространств [*1)]; $\mathcal{A}$ – оператор (правило), реализующий вычисления.

В первую очередь нас будут интересовать корректно поставлененые задачи вычисления.

Задача вычисления $y = \mathcal{A}(x)$ называется корректно поставленной, если для любых входных данных из некоторого класса решение задачи существует, единственно и устойчиво по входным данным (т.е. непрерывно зависит от входных данных задачи).

[*1)] Если не оговорено особо, то $\mathcal{Y}, \mathcal{X}$ – как правило *линейные, нормированные, полные,* т.е. *банаховы* пространства.

В сформулированное понятии *корректности* поставленной задачи (по Адамару) учтены достаточно естественные требования, действительно: чтобы численно решать задачу нужно быть уверенным, что её решение *существует*. Столь же разумны для конкретных условий и требования *единственности* решения, и, поскольку наши действия носят принципиально приближенный характер, то необходимо требование *устойчивости* решения.

Сделаем несколько замечаний об *устойчивости*. Нас интересует решение y задачи (1) соответствующее входным данным x. Реально мы имеем возмущенные входные данные с погрешностью δx, т.е. $x + \delta x$ и находим возмущенное решение

$$y + \delta y = \mathcal{A}(x + \delta x).$$

Эта погрешность входных данных порождает *неустранимую* погрешность решения

$$\delta y = \mathcal{A}(x + \delta x) - \mathcal{A}(x).$$

Если решение непрерывно зависит от входных данных, то

$$\|\delta y\| \to 0 \quad \text{всегда при} \quad \|\delta x\| \to 0$$

и задача (1) устойчива по входным данным.
Отсутствие устойчивости означает, что даже "небольшим" погрешностям δx могут соответствовать "большие" погрешности δy, т.е. построенное при расчёте решение будет сильно отличаться от истинного.

Применять непосредственно к такой неустойчивой задаче численные методы бессмысленно. Однако и не всякую формально устойчивую задачу удобно решать практически. Пусть имеет место оценка

$$\|\delta y\| \le C \cdot \|\delta x\|, \qquad \text{но} \qquad C\text{– велико.}$$

Задача формально устойчива, но *неустранимая ошибка* решения может быть большой. Это случай *плохой обусловленности* или *слабой устойчивости* задачи вычисления.

Приведем несколько примеров постановки *задачи вычисления* (1).

2.2 Примеры постановки задачи вычисления

1°.Задача нахождения корней полинома. Рассмотрим некоторый полином степени n в приведенном виде (старший коэффициент равен единице):

$$p_n(x) \equiv x^n + a_1 x^{n-1} + a_2 x^{n-2} + \cdots + a_{n-1} x + a_n$$

вообще говоря с комплексными коэффициентами $(a_k, x_n \in C)$.
Требуется определить его корни. Пусть E^n — n–мерное комплексное евклидово пространство. Положим, что компоненты некоторого вектора $\vec{z} = \{z_1, z_2, \ldots, z_n\}$ этого пространства являются корнями полинома $p_n(x)$, т.е.

$$p_n(z_i) = 0, \quad i = \overline{1, n}.$$

Тогда, в силу теоремы Безу, мы можем $p_n(x)$ записать в виде:

$$p_n(x) = x^n + a_1 x^{n-1} + a_2 x^{n-2} + \cdots + a_{n-1}x + a_0 = (x - z_1)(x - z_2)\cdots(x - z_n) = \prod_{i=1}^{n}(x - z_i).$$

Отсюда мы получаем известные формулы Виетта:

$$a_k = (-1)^k \sigma_k, \quad k = \overline{1, n}. \qquad (*)$$

Здесь σ_k — элементарные, симметричные относительно $z_1, z_2, \ldots, z_n$ однородные функции k–го порядка

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

(каждое σ_k содержит C_n^k слагаемых).

Таким образом формулы Виетта (*) сопоставляют каждому вектору $z \in E^n$ вектор $\vec{a} = \{a_1, a_2, \ldots, a_n\} \in E^n$ того же пространства, т.е. определяют отображение $\mathcal{V} : E^n \Rightarrow E^n$ пространства E^n *на себя.* С помощью этого отображения $\mathcal{V}$ задача определения корней полинома $p_n(x)$ формулируется следующим образом:
Для заданного вектора $\vec{a}$ *найти вектор* $\vec{z} \in E^n$ *такой, что*

$$\mathcal{V}(\vec{z}) = \vec{a}. \qquad (2)$$

В курсе высшей алгебры показано, что отображение $\mathcal{V}$ *взаимнооднозначное* и *взаимнонепрерывное*, т.е. задача (2) *корректна.*

2°.Основная задача линейной алгебры. Пусть дана матрица $A_{(p\times q)} = \|a_j^i\|_q^p$ и два евклидовых пространства E^p и E^q. Тогда определено отображение

$$A: \quad E^q \Rightarrow E^p; \quad \vec{y} = A\vec{x}, \quad \vec{y} \in E^p, \vec{x} \in E^q,$$

($\vec{x}, \vec{y}$ — столбцы соответствующих размерностей).
Основная задача линейной алгебры состоит в том, чтобы по заданному вектору $\vec{f} \in E^p$ найти вектор $\vec{x} \in E^q$ такой, что

$$A\vec{x} = \vec{f}. \qquad (3)$$

Задача (3) представляет собой задачу решения *системы линейных алгебраических уравнений* — СЛАУ. Связанная с решением СЛАУ ситуация нами подробно изучена в курсе *линейной алгебры*:

1) если $p = q$ и $det\, A \neq 0$, то задача (3) поставлена *корректно* (её решение дается формулами Крамера);

2) в остальных случаях, если система (3) совместна ($rang A = rang A'$), то решение неединственно. В противном случае решение вовсе отсутствует, т.е. задача (3) в этих случаях *некорректно поставлена.*

3°.Задача Коши для обыкновенного дифференциального уравнения.
Пусть требуется найти решение обыкновенного дифференциального уравнения (ОДУ), отвечающее начальному условию $y(a) = b$

$$\begin{cases} \frac{dy}{dx} = f(x,y), \quad a < x \le c \\ y(a) = b. \end{cases} \qquad (*)$$

Здесь a, b — заданные числа; $f(x,y)$ — определена в полосе $\Pi = \{(x,y); \quad a \le x \le c; \quad y \in (-\infty;\infty)\}$ и удовлетворяет в Π условиям теоремы о продолжимости решения (*) на отрезок $[a;c]$.

Обозначим через $\mathcal{R}_0$ множество всевозможных решений задачи Коши (*), отвечающих различным значениям начального условия b. Определим отображение $\mathcal{K}: \quad \mathcal{R}_0 \Rightarrow R^1$, полагая

$$\mathcal{K}(y(x)) = y(a), \quad \forall y \in \mathcal{R}_0.$$

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

$$\mathcal{K}(y(x)) = b. \qquad (4)$$

В курсе *дифференциальных уравнений* доказана *корректность* задачи (4).

Число рассмотренных примеров *задачи вычисления*можно было бы множить, но мы ограничимся рассмотренными примерами постановки *задачи вычисления.*

§3. Численное решение корректных задач Структура погрешности решения

3.1 Задача "вычисления". Погрешности

Обратимся снова к *задаче вычисления* (1)

$$y = \mathcal{A}(x).$$

В рассмотренных примерах (2)–(4) соответствующее *правило* $\mathcal{A}$ реализующее "вычисление" задано явно неконструктивно. Речь идёт по сути об обращении операторов $\mathcal{V}^{-1}, A^{-1}, \mathcal{K}^{-1}$, точнее о численной реализации обратного отображения для (2)–(4). Такая ситуация типична и лишний раз показывает, что, как правило, *вычисление* $\mathcal{A}$ не может быть "просто " реализовано. Чтобы преодолеть эти сложности задачу (1) заменяют другой, "близкой" к ней задачей, но уже которая "легко" решается численно. При этом в первую очередь анализируют вопрос о вносимых в решение погрешностях.

Есть четыре основных источника погрешности результата вычислений: *математическая модель*; *исходные данные* задачи; *приближнный метод* и *погрешность при реализации вычислений* (в частности *погрешность округления*):

$\delta_1 y$ *–погрешность математической модели*, связана с физическими допущениями при выборе математической модели и на анализе этой погрешности мы останавливаться не будем;

$\delta_2 y$ –*погрешность исходных данных*, порождает *неустранимую погрешность решения*

$$\delta_2 y = \mathcal{A}(x + \delta x) - \mathcal{A}(x);$$

$\delta_3 y$ –*погрешность метода.* Выражение $\mathcal{A}(x)$, вообще говоря, не может быть "просто" численно реализовано. Задачу $y = \mathcal{A}(x)$ заменяют "близкой" задачей

$$\bar{y} = \bar{\mathcal{A}}(\bar{x}), \tag{1'}$$

Мы переходим к другим функциональным пространствам $\mathcal{X}, \mathcal{Y} \Rightarrow \bar{\mathcal{X}}, \bar{\mathcal{Y}}$ элементы которых допускают сравнительно "простую" численную реализацию. Соответствующим образом меняется и отображение $\mathcal{A} \Rightarrow \bar{\mathcal{A}}$.

При этом естественно требовать, чтобы задача (1′) была *корректна* и чтобы решение $\bar{y}$ было близко к решению y. Величина

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

и представляет собой *погрешность метода.*

$\delta_4 y$ –*вычислительная погрешность.* При численной реализации $\bar{y}$, которая уже, по предположению, возможна получают элемент $\tilde{y}$, поскольку промежуточные результаты округлялись и т.п. Таким образом *вычислительная погрешность метода* может быть записана в виде

$$\delta_4 y = \bar{y} - \tilde{y} = \bar{\mathcal{A}}(\bar{x}) - \tilde{y}.$$

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

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

1) При проведении вычислений нужно стремиться, чтобы погрешность метода $\delta_3 y$ была бы в несколько раз меньше *неустранимой погрешности* решения $\delta_2 y$:

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

Теперь мы можем ещё раз очертить круг вопросов, рассматриваемых в рамках нашего лекционного курса "Численных методов" — это [*1)]

1) конструирование *дискретной* (или *вычислительной*) модели $\{\bar{\mathcal{X}}, \bar{x}, \bar{\mathcal{A}}\}$:

2) разработка на её основе соответствующих алгоритмов решения редуцированной задачи вычисления

$$\bar{y} = \bar{\mathcal{A}}(\bar{x});$$

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

[*1)] Предмет лекционного курса мог бы быть и более содержательным и обширным, но, как всегда, здесь есть свои, не зависящие от нашего желания, ограничения, определяемые спецификой учебного плана факультета.

3.2 Погрешность округления на t–разрядной ЭВМ

Остановимся несколько подробнее в рамках этого параграфа, но кратко, на анализе *вычислительной погрешности* $\delta_4 y$, обязанной погрешностям округления при реализации численного алгоритма.

1°.Погрешность единичного округления. 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

$$a = \operatorname{sign} a(a_n a_{n-1} \dots a_1 a_0, a_{-1} a_{-2} \dots)_r = \operatorname{sign} a(a_n r^n + a_{n-1} r^{n-1} + \dots + a_1 r + a_0 + \frac{a_{-1}}{r} + \frac{a_{-2}}{r^2} + \dots),$$

то такую форму записи числа a называют *представлением с фиксированной запятой*. Здесь $a_k \in \{0; 1; \dots; "(r-1)"\}$ — r–ичные цифры.
Представление числа a в форме с *плавающей запятой* или *нормализованное представление* означает его запись в виде

$$a = \operatorname{sign} a M r^p = \operatorname{sign} a \cdot r^p \cdot (\frac{b_1}{r} + \frac{b_2}{r^2} + \dots),$$

где p — порядок числа (целое); M — мантисса числа a, причем $1/r \le M < 1$, т.е. превая r–ичная цифра в записи мантиссы b_1 неравна нулю.

В современных ЭВМ в качестве основания системы счисления r выбирается двойка — $r = 2$. Тогда, если для записи мантиссы отводится только t двоичных разрядов, то это позволяет из диапазона $[M_0; M_\infty = {M_0}^{-1}]$ (для положительных чисел) записать лишь конечное число рациональных чисел, а все остальные вещественные числа подвергаются округлению при их представлении в ЭВМ.
Точность представления числа a с помощью округлённого числа $\tilde{a}$ характеризуется относительной погрешностью округления

$$\delta_a = \frac{|a - \tilde{a}|}{|a|}.$$

При простейшем способе округления *усечением*, когда все лишние разряды мантиссы просто отбрасываются, можно легко получить оценку величины относительной погрешности δ_a единичного округления. Действительно [*1)]

$$|a - \tilde{a}| = 2^p \left| \frac{b_{t+1}}{2^{t+1}} + \dots \right| \le 2^p \cdot \frac{1}{2^{t+1}} \left(1 + \frac{1}{2} + \dots\right) = 2^{p-t}.$$

С другой стороны[*2)] $|a| \ge 2^p \cdot (1/2)$. Таким образом для погрешности единичного округления получаем

$$\delta_a = \frac{|a - \tilde{a}|}{|a|} \le \frac{2^{p-t}}{2^{p-1}} = 2^{-(t-1)}.$$

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

$$\delta_a = 2^{-t} \equiv \varepsilon_M. \tag{5}$$

[*1)] Здесь при оценке все двоичные цифры в остатке заменены единицей $b_k \le 1, k \ge t+1$.
[*2)] Мы полагаем $a_i = 0$, при $i \ge 2$; $\quad a_1 = 1$ всегда.

Относительная погрешность представления числа с плавающей запятой в ЭВМ определяется числом разрядов мантиссы и не превышает машинного эпсилон $\varepsilon_M = 2^{-t} (\sim 10^{-12})$.

Опираясь на оценку (5) мы можем считать, что само число a и его округлённое значение $\tilde{a}$ связаны соотношением

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

где $|\varepsilon_a| \le \varepsilon_M = 2^{-t}$. Однако отметим, что для чисел $|a| < M_0$ в результате округления получим $\tilde{a} = 0$ и тем самым для этих чисел $\varepsilon_a = -1$ (!).

Арифметическое Устройство (АУ) современных ЭВМ сконструировано таким образом, что любая арифметическая операция при последующем округлении даёт относительную ошибку не более ε_M.

Для оценки влияния погрешности округлений на результат того или иного вычислительного алгоритма пользуются предположением о том, что *результат вычислений, искажнный погрешностью округления совпадает с результатом точного вычисления по тому же алгоритму, но с иными — $\tilde{x}$, входными данными.*

Таким образом

$$\tilde{y} = \bar{\mathcal{A}}(\tilde{x}) \quad \text{и} \quad \delta_4 y = \bar{y} - \tilde{y} = \bar{\mathcal{A}}(\bar{x}) - \bar{\mathcal{A}}(\tilde{x}).$$

Это допущение позволяет связать анализ *вычислительной погрешности* $\delta_4 y$ с анализом *устойчивости* алгоритма $\bar{\mathcal{A}}$ по *входным данным.* Ограничимся рассмотрением

Пример. Рассмотрим задачу о нахождении произведения n сомножителей

$$z_n = \prod_{k=1}^{n} y_k.$$

Пусть вычисления реализованы по алгоритму $\bar{\mathcal{A}}$ следующим образом:

$$\begin{cases} z_k &= y_k \cdot z_{k-1}, \quad k = 1, 2, \ldots, n \\ z_0 &= 1. \end{cases}$$

Предположим, что в результате округлений вместо точного значения z_{k-1} получено знадение $\tilde{z}_{k-1}$. Тогда вместо величины $y_k \tilde{z}_{k-1}$ получим величину

$$\tilde{z}_k = \mathrm{fl}\,(y_k \cdot \tilde{z}_{k-1}) = y_k \cdot \tilde{z}_{k-1}(1 + \varepsilon_k); \quad |\varepsilon_k| \le \varepsilon_M.$$

Таким образом мы получили алгоритм $\tilde{\mathcal{A}}$[*1)]

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

Оценим результирующую относительную погрешность

$$\delta_{z_n} = \left| \frac{z_n - \tilde{z}_n}{z_n} \right| = \frac{\left| \prod\limits_{k=1}^{n} y_k - \prod\limits_{k=1}^{n} (1 + \varepsilon_k) y_k \right|}{\left| \prod\limits_{k=1}^{n} y_k \right|} \le (1 + \varepsilon_M)^n - 1 = n\varepsilon_M + O(\varepsilon_M^2).$$

или, пренебрегая слагаемыми второго и больших порядков по ε_M получим окончательно

$$\delta_{z_n} \le n\varepsilon_M.$$

[*1)] Структура полученного алгоритма $\bar{\mathcal{A}}$ подтверждает сформулированное допущение.

ГЛАВА II

ЗАДАЧА ИНТЕРПОЛЯЦИИ и ПРИБЛИЖЕНИЯ ФУНКЦИЙ

§1. Постановка задачи интерполяции функции

1.1 Постановка задачи интерполяции

Мы ограничимся рассмотрением задачи интерполяции для функции одной вещественной переменной.
Пусть на отрезке $a \le x \le b$ задана невырожденная сетка $\overline{\omega}$

$$\overline{\omega} = \{x_0 = a < x_1 < x_2 < \ldots < x_n = b\}$$

и в её узлах известны значения функции $y = f(x)$:

$$f(x_0) \equiv y_0 \quad ; \quad f(x_1) \equiv y_1 \quad ; \quad \cdots \quad f(x_{n-1}) \equiv y_{n-1} \quad ; \quad f(x_n) \equiv y_n.$$

В задаче интерполяции требуется построить интерполяционную функцию — *интерполянту* $g(x)$ так, чтобы её значения совпадали бы со значениями интерполируемой функции $f(x)$ в узлах сетки $\overline{\omega}$, т.е.

$$g(x_i) = y_i \quad ; \quad i = \overline{0, n} \tag{1}$$

При этом основная цель интерполяции получить *быстрый* и *экономичный* алгоритм вычисления значений функции $g(x)$ для значений x не содержащихся в исходной таблице, т.е. $x \in [a; b]$ и $x \neq x_i$.

Как выбрать интерполянту $g(x)$ и как оценить погрешность интерполяции $\|f(x) - g(x)\|$ (в некоторой норме)?

Отвечая на эти вопросы, мы ограничимся рассмотрением задачи *линейной интерполяции* (относительно базисных функций $\Phi_k(x)$), т.е. рассмотрим ситуацию, когда интерполирующая функция $g(x)$ строится в виде линейной комбинации некоторых базисных функций $\{\Phi_k(x)\}$ в свою очередь достаточно элементарных

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

где $\{\Phi_k(x)\}$ — фиксированные, линейно независимые функции; $c_0, c_1, \ldots, c_n$ — неопределенные пока коэффициенты.

Из условий интерполяции (1) мы получаем систему $(n + 1)$ уравнения относительно коэффициентов $\{c_k\}$

$$\sum_{k=0}^{n} c_k \Phi_k(x_i) = y_i \quad , \quad i = 0, \ldots, n.$$

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

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

Такую систему интерполяционных функций называют *чебышевской системой интерполяционных функций.* В этом случае на данной сетке $\overline{\omega}$ по известным значения $y_i, i=0,\dots,n$ однозначно определяются коэффициенты интерполяционного многочлена $\{c_k\}_{k=\overline{0,\dots,n}}$.

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

В качестве систем линейно независимых интерполяционных функций $\{\Phi_k(x)\}$ чаще всего выбирают степенные или полиномиальные функции $\Phi_k(x)=x^k$; тригонометрические функции $\Phi_k(x)=\{{}^{\cos(kx)}_{\sin(kx)}\}$; показательные функции $\Phi_k(x)=e^{kx}$ и другие.

Мы ограничимся рассмотрением случая *полиномиальной* интерполяции.

Замечания: Табличный способ задания функции $y=f(x)$ на сетке $x\in\overline{\omega}$ и связанная с этим необходимость интерполяции функции наиболее характерны для представления результатов *физического* эксперимента и для описания *дискретной* или *вычислительной* модели на ЭВМ.

Задача интерполяции, как задача перестройки таблиц значений функции с одной сетки на другую, является характерной задачей обработки данных *физического* эксперимента.

В дальнейшем мы используем решение задачи интерполяции при построении приближенных методов вычисления интегралов; при разностной аппроксимации дифференциальных уравнений на основе интегральных тождеств; в задачах минимизации и в других вопросах.

Постановка задачи интерполяци в форме (1) не единственно возможная. Возможны и другие постановки *задачи интерполяции*, например задача интерполяции Эрмита [*1)] и другие виды задачи интерполяции.

Итак рассмотрим

§2. Полиномиальная интерполяция

2.1 Существование и единственность интерполяционного полинома

Пусть входной информацией для нас является множество точек $\{(x_k;y_k)\}_{k=\overline{0,\dots,n}}$. Интерполянту $g(x)$ мы будем искать в виде *интерполяционного полинома*

$$g(x)\equiv P_n(x)=\sum_{k=0}^{n}c_kx^k. \tag{2}$$

[*1)] Эта постановка связана с интерполяцией функции по известным значения функции и её производных до n-ого порядка включительно, заданными в некоторой точке $x=x_0$.

Условия интерполяции (1) $g(x_i) = y_i$ приводят к системе линейных алгебраических уравнений для коэффициентов $\{c_k\}$:

$$\left\{ \begin{array}{rcl} c_0 + c_1 x_0 + \cdots + c_n x_0^n &=& y_0 \\ c_0 + c_1 x_1 + \cdots + c_n x_1^n &=& y_1 \\ \cdots & & \\ c_0 + c_1 x_n + \cdots + c_n x_n^n &=& y_n. \end{array} \right.$$

Определитель этой системы (определитель Вандермонда) отличен от нуля на произвольной невырожденной сетке $\overline{\omega}$

$$\triangle = \left| \begin{array}{ccccc} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \ldots & & & & \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{array} \right| =$$

$$= (x_n - x_{n-1}) \ldots (x_n - x_0)(x_{n-1} - x_{n-2}) \ldots (x_{n-2} - x_0) \ldots (x_1 - x_0) =$$

$$= \prod_{0 \le m < k \le n} (x_k - x_m) \neq 0.$$

Тем самым, система функций $\{x^k\}$ — *чебышевская* система интерполяционных функций на $[a; b]$ и справедлива

Теорема. *Интерполяционный полином (2) существует и единственен.* [*1)]

2.2 Интерполяционный полином Лагранжа

При построении интерполяционного полинома $P_n(x)$ (2) 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*, $\{l_k^{(n)}(x)\}_{k=0,\ldots,n}$, определенная из соображений: каждый $l_k^{(n)}(x)$ - полином n-ой степени, равный нулю во всех узлах сетки $\overline{w}$ кроме k-го, где он равен 1

$$l_k^{(n)}(x_i) = \left\{ \begin{array}{ll} 1, & i = k \\ 0, & i \neq k \end{array} \right. \qquad i, k = \overline{0, n}$$

Нетрудно построить эти полиномы. Действительно, зная корни полинома мы можем утверждать, что полином

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

решает поставленную задачу.

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

$$w(x) \equiv w_{\overline{0,n}}(x) = (x - x_0) \ldots (x - x_n) = \prod_{i=0}^{n} (x - x_i)$$

[*1)] Однако *интерполяционный полином* может быть записан в различной форме.

Найдем его производную в точке $x = x_k$, имеем

$$w'(x)|_{x=x_k} = \left|\begin{array}{l}(n+1) \text{ слагаемое, но все, содер-}\\ \text{жащие скобку } (x - x_k) \text{ при под-}\\ \text{становке } x = x_k \text{ дадут } 0\end{array}\right| = (x_k - x_0)(x_k - x_1)\dots(x_k - x_n).$$

Тогда

$$l_k(x) = \frac{w(x)}{(x - x_k)w'(x_k)}$$

и есть *базис полиномов Лагранжа.*

Отметим, что *построенный базис единственен.* Действительно, если существует полином $\bar{l}_k(x)$ при тех же условиях, то полином

$$q^{(n)}(x) = l_k(x) - \bar{l}_k(x)$$

есть полином n-ой степени обращающийся в ноль в $(n+1)$ -ой точке $x_0, \dots, x_n$. Он тождественно равен нулю и, следовательно, $\bar{l}_k(x) = l_k(x)$.

Теперь легко записать решение задачи полиномиальной интерполяции. Полином $y_k \cdot l_k(x)$ принимает в узле x_k значение y_k и равен нулю во всех осталных узлах сетки $\overline{w}$ (при $x_i \neq x_k$). Тогда

$$L_n(x) \equiv \sum_{k=0}^{n} l_k(x) y_k = \sum_{k=0}^{n} f(x_k) \frac{w(x)}{(x - x_k)w'(x_k)} \tag{3}$$

представляет собой полином степени не выше n и $L_n(x_i) = y_i$, т. е. является *интерполяционным полиномом.*

Формулу (3) называют *интерполяционным формулой Лагранжа*, а соответствующий полином $L_n(x)$ — *интерполяционным полиномом Лагранжа.*

Замечания. Нетрудно оценить число арифметических действий при вычислении по формуле (3). В главном порядке по n это есть величина $O(n^2)$.

2.3 Интерполяционный полином Ньютона

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

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

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

$$f(x_i, x_j) = \frac{f(x_i) - f(x_j)}{x_i - x_j} = \left|\begin{array}{l}\text{они}\\ \text{симмет-}\\ \text{ричны}\end{array}\right| = \frac{f(x_j) - f_i}{x_j - x_i} = f(x_j, x_i).$$

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

$$f(x_i, x_j, x_k) = \frac{f(x_i, x_j) - f(x_j, x_k)}{x_i - x_k}; \qquad \text{и т. д.}$$

Если известны *разделенные разности* k -го порядка, то *разделенные разности* $(k+1)$-го порядка на узлах $x_j, x_{j+1}, \ldots, x_{j+k+1}$ определяются как

$$f(x_j, x_{j+1}, \ldots, x_{j+k+1}) = \frac{f(x_j,\ldots,x_{j+k}) - f(x_{j+1},\ldots,x_{j+k+1})}{x_j - x_{j+k+1}} = \frac{f(x_{j+1},\ldots,x_{j+k+1}) - f(x_j,\ldots,x_{j+k})}{x_{j+k+1} - x_j}.$$

Заметим, что при рассмотрении введенных понятий соседство узлов сетки не обязательно, важно их количество: для *разделенной разности* k-го порядка — $(k+1)$ узел.

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

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

Отметим особенности этих разделенных разностей:

Если табличная функция сама по себе полином n-ой степени, т. е. $f(x) \equiv P_n(x) \equiv p(x)$, то её 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, 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 $x = x_1$.

Таким образом $(n+1)$ *разделенная разность* для полинома n-ой степени *тождественно равна нулю.*

Из определения разделенных разностей получим

$$
\begin{aligned}
&p(x) = p(x_0) + (x - x_0)p(x, x_0)\\
&p(x, x_0) = p(x_0, x_1) + (x - x_1)p(x, x_0, x_1)\\
&p(x, x_0, x_1) = p(x_0, x_1, x_2) + (x - x_2)p(x, x_0, x_1, x_2)\\
&\dots\\
&p(x, x_0, x_1, \dots, x_{n-2}) = p(x_0, x_1, \dots, x_{n-1}) + (x - x_{n-1})\, p(x, x_0, x_1, \dots, x_{n-1})\\
&p(x, x_0, x_1, \dots, x_{n-1}) = p(x_0, x_1, \dots, x_n) + (x - x_n)\underbrace{p(x, x_0, x_1, \dots, x_n)}_{\equiv 0}.
\end{aligned}
$$

Осуществляя обратную подстановку (рекурентно) найдем

$$
\begin{aligned}
p(x) = p(x_0) + (x - x_0)\Big\{ & p(x_0, x_1) + (x - x_1)\{p(x_0, x_1, x_2) + \dots\\
\dots + (x - x_{n-2})[p(x_0, \dots, x_{n-1}) + (x - x_{n-1})(p(x_0, \dots, x_n) + 0]\} \dots & \Big\} =\\
= p(x_0) + (x - x_0)p(x_0, x_1) + (x - x_0)(x - x_1)p(x_0, x_1, x_2) + \dots\\
\dots + (x - x_0)(x - x_1)\dots(x - x_{n-2})p(x_0, \dots, x_{n-1}) +\\
+(x - x_0)(x - x_1)\dots(x - x_{n-1})p(x_0, \dots, x_n).
\end{aligned}
$$

Осталось учесть последнее: если $p(x)$ - интерполяционнный полином для функции $f(x)$, то $p(x_i) = f(x_i)$ и мы получим явную форму записи интерполяционного многочлена:

$$
\begin{aligned}
N_n(x) \equiv f(x_0) + (x - x_0) f(x_0, x_1) + (x - x_0)(x - x_1) f(x_0, x_1, x_2) + \cdots\\
+ (x - x_0) \dots (x - x_{n-1}) f(x_0, x_1, \dots x_n)
\end{aligned}
\tag{4}
$$

в виде *интерполяционного многочлена Ньютона.*

2.4 Погрешность полиномиальной интерполяции

Остановимся теперь на вопросе о *погрешности* полиномиальной интерполяции. Пусть $P_n(x)$ обозначает полином степени не выше, чем n, который решает задачу интерполяции функции $f(x)$ на сетке $\overline{\omega}$ (при этом мы отвлекаемся от способа получения этого полинома в виде (3) или (4)).

Требуется, в некотором смысле, оценить разность

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

Чтобы провести эту оценку через $f(x)$, предположим, что наша функция $f(x)$ имеет непрерывные до $(n+1)$ порядка включительно производные на $[a; b]$, то есть $f(x) \in C^{(n+1)}[a; b]$. Рассмотрим вспомогательную функцию

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

где $A = const$; $\omega(z) = \omega_{\overline{0,n}}(z) = \prod\limits_{i=0}^{n}(z - x_i)$ - введенный ранее многочлен $(n+1)$ степени.

И определим const A из того условия, чтобы в произвольной фиксированной точке $x \neq x_k$; $k = \overline{0, n}$ выполнялось равенство $\varphi(x) = 0$. Тогда

$$A = \left.\frac{f(z) - P_n(z)}{\omega(z)}\right|_{z=x} = \frac{f(x) - P_n(x)}{\omega(x)}, \qquad \text{ибо } \omega(x) \neq 0 \quad \text{при } x \neq x_k.$$

С другой стороны, определенная так функция $\varphi(z)$ - непрерывна на $\left[\widetilde{a}=\min(x,x_0,x_1,...x_n);\widetilde{b}=\max(x,x_0,x_1,...x_n)\right]$, имеет производную и обращается в нуль в $(n+2)$ точках $x,x_0,x_1,...x_n$ отрезка $\left[\tilde{a};\tilde{b}\right]$. По теореме *Ролля*, отсюда следует, что существует $(n+1)$ внутренняя точка на $\left[\tilde{a};\tilde{b}\right]$, где $\varphi'(z)=0$. Аналогично, существует n точек, где $\varphi''(z)=0$ и т.д.
Следовательно, существует одна точка, в которой

$$\varphi^{(n+1)}(z)=0,$$

то есть $\exists\xi\in\left(\widetilde{a};\widetilde{b}\right)$ такая, что $\varphi^{(n+1)}(\xi)=0$. Но

$$\varphi^{(n+1)}(z)=f^{(n+1)}(z)-A(n+1)!\,|_{z=\xi},$$

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

$$A=\frac{f^{(n+1)}(\xi)}{(n+1)!}=\frac{f(x)-P_n(x)}{\omega(x)}.$$

Для представления остаточного члена $R_n(x)$, получаем фомулу:

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

здесь ξ -формально зависит от x[*1)].
Это и есть исходное выражение, которое позволяет получить оценку погрешности интерполяции. Имеем

$$|f(x)-P_n(x)|\le\frac{M_{n+1}}{(n+1)!}|\omega(x)|, \tag{5}$$

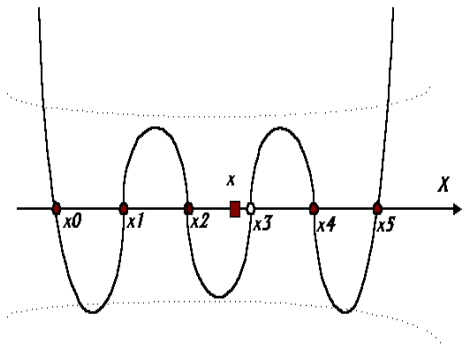
где $M_{n+1}=\max\limits_{\left[\widetilde{a};\widetilde{b}\right]}\left|f^{(n+1)}(x)\right|$[*2)].

Дальнейшее использование оценки (5) 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 $\overline{\omega}$ с постоянным шагом $h=\frac{b-a}{n}$:

2) Узлы интерполяции на этой сетке выбраны подряд.

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



Многочлен

$$\omega(x)=(x-x_0)(x-x_1)...(x-x_4)(x-x_5)$$

полином шестой степени.

Вблизи центральных узлов интерполяции экстремум $|\omega(x)|$ невелик. Для крайних интервалов — побольше.

Вне сетки узлов $|\omega(x)|$ быстро возрастает.

[*1)] От x зависит A в представлении погрешности.
[*2)] Поскольку $f^{(n+1)}(x)$ непрерывна.

Рассмотенный эскиз частично обосновывает вывод:

1)*экстраполяция* ненадежна. Результатам, если $x \notin [a;b]$ нельзя доверять;

2)при интерполяции на равномерной сетке выгодно так выбирать узлы $\{x_i\}$ из таблицы, чтобы точка x была по возможности близка к центру конфигурации узлов. Это обеспечивает большую точность и надежность интерполяции.

Сравнительно просто дальнейшая оценка погрешности интерполяции проводится в случае *нечетного* $n = 2k+1$ (когда на сетке $\overline{\omega}$ расположены $2k+2$ узла и имеется $2k+1$ интервал длины h).

Пусть при этом рассматриваемое x находится в центральном интервале $x \in (x_k; x_{k+1})$. На этом интервале экстремум $\omega(x)$ (в силу симметрии $\omega(x)$ относительно точки $x_0 + kh + \frac{h}{2}$) достигается точно в середине $(k+1)$-го интервала сетки $\overline{\omega}$, и его можно оценить:

$$\left|\omega\left(x_0 + kh + \tfrac{h}{2}\right)\right| = \left(\left(kh + \tfrac{h}{2}\right)\left((k-1)h + \tfrac{h}{2}\right)...\tfrac{h}{2}\right)^2 =$$

$$= \left[\frac{h^{k+1}(2k+1)(2k-1)...3\cdot 1}{2^{k+1}}\right]^2 = \left(\frac{h^{k+1}\cdot(2k+1)!}{2^{k+1}\cdot 2^k\cdot k!}\right)^2.$$

Далее, применяя формулу Стирлинга $n! \approx \sqrt{2\pi}\left(\frac{n}{e}\right)^n$ при известной аккуратности в неравенствах получаем окончательную оценку

$$|f(x) - P_n(x)| \le \sqrt{\frac{2}{\pi n}} M_{n+1} \left(\frac{h}{2}\right)^{n+1}; \quad x \in (x_k, x_{k+1}) \tag{6}$$

погрешности интерполяции в центральном интервале сетки $\overline{\omega}$.

Замечания:

1) Если a'priori известна оценка M_{n+1} для $max|f^{(n+1)}(x)|$ на $[a;b]$, то из формулы (6) можно найти число узлов $(n+1)$, необходимое для интерполяции с заданной точностью;

2) Из формулы (6) видно, что если перейти к интерполяции по таблице с более мелким шагом (при томже числе узлов сетки n), то погрешность интерполяции будет убывать как величина порядка $O(h^{n+1})$ (ассимптотика погрешности $\delta_3 y = O(h^{n+1})$). Поэтому говорят, что интерполяционный многочлен $P_n(x)$ обеспечивает $(n+1)$-ый порядок точности интерполяции и интерполяция имеет погрешность $O(h^{n+1})$.

2.5 Сходимость интерполяционного процесса

Остановимся несколько более подробно на сходимости интерполяционного процесса. Говоря о сходимости интерполяционного процесса, ищут ответ на вопрос о стремлении в некотором смысле к нулю погрешности интерполяции $\|f(x) - P_n(x)\|$ при неограниченном увеличении числа узлов интерполяции $n \to \infty$.

Для этого рассматривают на $[a;b]$ последовательность сеток [*1)]

$$\{\overline{\omega}(x)\} : \overline{\omega}_0 = \left\{x_0^{(0)}\right\}; \overline{\omega}_1 = \left\{x_0^{(1)}, x_1^{(1)}\right\}; ...\overline{\omega}_n = \left\{x_0^{(n)}, x_1^{(n)}, ...x_n^{(n)}\right\}; ...$$

[*1)] Таких последовательностей сеток бесконечно много.

Рассмотрим соответствующее им множество интерполяционных полиномов $\{P_n(x)\}$ (зависящих от $\{\overline{\omega}_n(x)\}$), построенных для интерполируемой функции $f(x)$ на сетке $\overline{\omega}_n = \left\{x_i^{(n)}\right\}_{\overline{0,n}}$.

Говоря о сходимости функциональной последовательности $\{P_n(x)\}$, обычно имеют в виду:

1)поточечную сходимость к $f(x)$ на $[a;b]$:

$\forall x, \lim\limits_{n\to\infty} P_n(x) = f(x)\, x \in [a;b]; \quad P_n(x) \to f(x)$ на $[a;b]$;

2)равномерную сходимость к $f(x)$ на $[a;b]$:

$\lim\limits_{n\to\infty} \sup\limits_{[a;b]} |f(x) - P_n(x)| = 0; \quad P_n(x) \Rightarrow f(x)$ на $[a;b]$;

3)сходимость к $f(x)$ в среднем с весом $\rho(x) \geq 0$ на $[a;b]$:

$\lim\limits_{n\to\infty} \int\limits_a^b (f(x) - P_n(x))^2 \rho(x)\, dx = 0; \quad P_n(x) \underset{[a;b]}{\overset{\text{ср}}{\to}} 0$ в седнем на $[a;b]$.

Подчеркнём ещё раз, что последовательность сеток $\overline{\omega}_n$ фиксирована при рассмотрении соответствующих пределов.

В нашем курсе мы ограничимся практическими рекомендациями.

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

1)либо сохраняя степень интерполяционного полиноми, уменьшать шаг сетки $(h \to 0, n = const)$;

2)либо сохраняя шаг сетки, увеличивать число используемых узлов интерполяции на $[a;b]$, то есть увеличивать степень интерполяционного многочлена: $h \to 0, n = const$.

1) Уменьшение шага сетки $(h \to 0)$. Если $f(x) \in C^{(n+1)}_{[a;b]}$, то, как мы уже отмечали, погрешность метода при интерполяции многочленом $P_n(x)$, есть, согласно (6), величина порядка $O(h^{n+1})$, т.е. $|f(x) - P_n(x)|$ неограниченно убывает при $h \to 0$ и при этом интерполяционный многочлен сходится к $f(x)$ в некотором смысле "равномерно".

Точнее говоря, для каждого $x \in [a;b]$ выбираются свои узлы интерполяции, ближайшие именно на данной сетке к точке x. При этом точка x лежит заведомо между крайними узлами, использованными при построении интерполяционного многочлена. Тогда, равномерно по x, можно в формуле (5) провести оценку

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

nh–длина отрезка $\left[\widetilde{a};\widetilde{b}\right]$ (своего для каждой сетки).

Отсюда, для заданной точности ε, можно получить условие на величину шага h сетки, обеспечивающего данную точность:

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

Напомним, что n фиксировано и тем самым

$$h \leq \sqrt[n+1]{\frac{\varepsilon\,(n+1)!}{M_{n+1}}} \cdot \frac{1}{n}.$$

Все сетки с данным и более мелким шагом для любой точки $x \in [a;b]$ *дают погрешность интерполяции многочленом* $P_n(x)$*, с указанным образом расположенными узлами, не боле, чем* ε*.*

2) Увеличение числа узлов $(n \to \infty)$**.**

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

а) не известно, как быстро растет оценка максимума модуля производной M_{n+1} с ростом её порядка;

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

В общем случае свойство сходимости или расходимости интерполяционного процесса зависит как от выбора последовательности сеток $\{\overline{\omega}_n\}$, так и от гладкости интерполируемой функции $f(x)$.

Ограничимся приведением отдельный фактов:

1) Легко привести примеры несложных функций, для которых интерполяционный процесс расходится. Так, для $y = |x|$, интерполяция на *равномерной* на $[-1;1]$ сетке не дает поточечной сходимости ни в одной точке x, кроме $x \in \{-1;0;1\}$

$$P_n(x) \not\to f(x); \quad \forall x \in [a;b] \setminus \{-1;0;1\};$$

2) Если $f(x)$ *целая* функция, то есть может быть разложена в степенной ряд с бесконечным $(R = \infty)$ радиусом сходимости, то при *произвольном (!)* выборе сеток $\{\overline{\omega}_n\}$

$$P_n(x) \rightrightarrows f(x) \quad \text{на } [a;b].$$

С этой точки зрения *целые* функции хороши, но их запас не столь "велик" для практических целей.

3) **Теорема** *(Фабера). Для любой последовательности сеток* $\{\overline{\omega}_n\}$ *найдётся непрерывная на* $[a;b]$ *функция* $f(x)$ *такая, что для неё нет равномерной сходимости* $P_n(x)$

$$P_n(x) \not\rightrightarrows f(x) \quad \text{на } [a;b];$$

С другой стороны

4) **Теорема** *(Марцинкевича). Для любой непрерывной на* $[a;b]$ *функции* $f(x) \in C[a;b]$ *найдется такая последовательность сеток* $\{\overline{\omega}_n\}$ *такая, что име—т место равномерная сходимость интерполяционных полиномов*

$$P_n \rightrightarrows f(x) \quad \textit{на } [a;b].$$

5) Сходимость интерполяционного многочлена в среднем $P_n(x) \overset{\text{ср}}{\to} f(x)$ на $[a;b]$ можно всегда обеспечить, выбирая на $[a;b]$ специальную сетку. Пусть $\{\Phi_n(x)\}$ - система ортогональных на $[a;b]$ с весом $\rho(x)$ полиномов; пусть $\left\{x_i^{(n)}\right\}_{i=\overline{1,n}}$ — нули этих полиномов (они все лежат внутри $(a;b)$ и с ростом n они перемежаются). Используя эти точки в качестве узлов интерполяции, можно утверждать, что

$$\lim_{n\to\infty} \int_a^b \left(f(x) - P_{n-1}(x)\right)^2 \rho(x)\, dx = 0.$$

(О подобного рода сетках мы специально поговорим в проблеме интегрирования.)
Общий вывод: В практике вычислений избегают использования интерполяционных полиномов высокой степени. Вместо этого для интерполяции $f(x)$ на большом отрезке используют *кусочно-полиномиальную интерполяцию.*

§3. Сплайн-интерполяция

Определение. *Сплайном порядка p на сетке $\overline{w}_n$ называется кусочно - полиномиальная порядка p функция, имеющая на $[a,b]$ непрерывные до $(p-1)$-го порядка включительно производные.*

Как мы постараемся показать преимуществом сплайнов перед обычной полиномиальной интерполяцией многочленом является, во-первых, их *сходимость*, и, во-вторых, *устойчивость* процесса их вычисления.

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

3.1 Определение кубического сплайна

Пусть на отрезке $[a,b]$ определена непрерывная (в дальнейшем достаточно гладкая) функция $f(x)$; задана невырожденная сетка $\overline{\omega}_n$

$$\overline{\omega}_n = \{a = x_0 < x_1 < x_2 < \ldots < x_n = b\}.$$

Обозначим значения $f(x)$ в узлах сетки через $y_i = f(x_i)$. Тогда
Определение. *Кубическим сплайном $s_3(x) \equiv s(x)$ на данной сетке $\overline{w_n}$ называется кусочно - полиномиальная 3-го порядка функция, удовлетворяющая следующим требованиям:*
1) На каждом частичном интервале $[x_{k-1}, x_k]$ многочлен $s(x)$ — многочлен 3-ей степени:

$$\begin{gathered} s(x) = a_k + b_k(x - x_k) + \tfrac{c_k}{2!}(x - x_k)^2 + \tfrac{d_k}{3!}(x - x_k)^3 \\ x_{k-1} \le x \le x_k, \quad k = 1, 2, \ldots, n \end{gathered} \tag{s1}$$

2) Функция $s(x)$, её первая и вторая производные непрерывны на $[a,b]$, т. е.

$$s^{(l)}(x - 0) = s^{(l)}(x + 0); \quad ; \forall x \in [a,b], \quad l = 0, 1, 2 \tag{s2}$$

(нужно проверять лишь в узлах сетки).
3) В узлах сетки $\overline{w_n}$ функция $s(x)$ удовлетворяет условиям интерполяции

$$s(x_i) = y_i; \qquad i = 0, \ldots, n \tag{s3}$$

4) Дополнительным, в некотором смысле естественным, условием для единственности определения сплайна является краевое условие в граничных точках сетки x_0 и x_n. Ограничимся рассмотрением случая нулевой кривизны сплайна $s(x)$. Тогда в этих точках

$$s''(x_0) = s''(x_n) = 0. \tag{s4}$$

3.2 Существование и единственность кубического интерполяционного сплайна $s(x)$

Теорема. *Сплайн $s(x)$ при условии $(s1 - s4)$ существует и единственен.*
Приведем конструктивное доказательства этого факта. Для сплайна $s(x)$ нам понадобятся производные до второго порядка включительно. Найдём

$$s'(x) = b_k + c_k(x - x_k) + \frac{d_k}{2!}(x - x_k)^2$$
$$s''(x) = c_k + d_k(x - x_k).$$

Причем в точке x_k имеем

$$s(x_k) = a_k$$
$$s'(x_k) = b_k$$
$$s''(x_k) = c_k.$$

Из определения интерполяционного сплайна $s(x)$ в узлах интерполяции должно иметь место

$$\begin{aligned} &s(x_i) = y_i, && i = \overline{0, n} \\ &s(x_i - 0) = s(x_i + 0), && i = \overline{1, n-1} \\ &s'(x_i - 0) = s'(x_i + 0), && i = \overline{1, n-1} \\ &s''(x_i - 0) = s''(x_i + 0), && i = \overline{1, n-1} \end{aligned}$$

Из условие интерполяции (s3) определим:

$$s(x_k) = a_k = y_k, \qquad k = 1, \ldots, n.$$

Положим $a_0 \equiv y_0$. Тем самым все коэффициенты $\{a_k\}$ определны.
Непрерывность производных сплайна $s(x)$ во внутренних узлах и найденные $\{a_k\}$ дают:
1) Непрерывность $s(x)$ в x_k:

$$a_k = a_{k+1} + b_{k+1}(x_k - x_{k+1}) + \frac{c_{k+1}}{2}(x_k - x_{k+1}) + \frac{d_{k+1}}{6}(x_k - x_{k+1})^3$$

где $k = 0, \ldots, n-1$.
Введем обозначения шага сетки $h_{k+1} = x_{k+1} - x_k$ (на $(k+1)$-ом интервале) и заменим $(k+1)$ на k. Тогда

$$b_k h_k - \frac{c_k}{2}h_k^2 + \frac{d_k}{6}h_k^3 = y_k - y_{k-1}, \qquad k = 1, \ldots, n; \tag{$*$}$$

2)Непрерывность $s'(x)$ в т. x_k даёт:

$$b_k = b_{k+1} + c_{k+1}(x_k - x_{k+1}) + \frac{d_{k+1}}{2}(x_k - x_{k+1})^2 \qquad k = 1, \ldots, n-1$$

что нетрудно, аналогично рассмотренному выше, преобразовать к виду

$$c_k h_k - \frac{d_k}{2}h_k^2 = b_k - b_{k-1} \qquad k = 2, \ldots, n \tag{$**$}$$

3)Из непрерывности $s''(x)$ в т. x_k найдем:

$$\begin{aligned} c_k = c_{k+1} + d_{k+1}(x_k - x_{k+1}) \quad &\text{или} \\ d_k h_k = c_k - c_{k-1}, \quad & \end{aligned} \tag{$***$}$$

где $k = 2, \ldots, n$.
Добавим к (3n-2) уравнениям (*), (**) и (***) граничные условия (s4)

$$\begin{array}{lll} s''(x_0) = 0 & \leftrightarrow & c_1 + d_1(x_0 - x_1) = 0 \quad \leftrightarrow \quad d_1 h_1 = c_1 - c_0; \\ s''(x_n) = 0 & \leftrightarrow & c_n = 0. \end{array}$$

Если положить $c_0 \equiv 0$, то граничные условия в точке x_0 в точности дают уравнение (**) при $k = 1$. Итак для определения коэффициентов $\{a, b, c\}$ сплайна $s(x)$ мы получим систему уравнений

$$\begin{cases} h_k b_k - \frac{c_k}{2} h_k^2 + \frac{d_k}{6} h_k^3 = y_k - y_{k-1}, & k = \overline{1,n} \Rightarrow b_k; \quad b_{k-1} & (1*) \\ c_k h_k - \frac{d_k}{2} h_k^2 = b_k - b_{k-1}, & k = \overline{2,n} & (2*) \\ h_h d_k = c_k - c_{k-1}, & k = \overline{2,n} \Rightarrow d_k; \quad d_{k-1} & (3*) \\ c_0 = 0, \quad c_n = 0 & & \end{cases}$$

Дальнейшее упрощение полученной системы связано с явным выражением b_k, b_{k-1}, из (1*) и d_k, d_{k-1}, из (3*)и подстановкой найденных выражений в (2*). Тогда

$$\begin{cases} c_0 = 0 \\ c_{k-1} h_k + 2(h_k + h_{k+1}) c_k + c_{k+1} h_{k+1} = 6 \left(\frac{y_{k+1} - y_k}{h_{k+1}} - \frac{y_k - y_{k-1}}{h_k} \right) \\ \qquad\qquad k = 1, \ldots, n-1 \\ c_n = 0 \end{cases} \tag{7}$$

получаем систему линейных алгебраических уравнений — СЛАУ с трехдиагональной матрицей.

- В силу диагонального преобладания элементов матрицы системы (7) её решение существует и единственно[*1)];
- Решения (7) эффективно строится методом *прогонки*;
- По найденным коэффициентам $\{c_k\}$ из (*) явно находятся $\{b_k\}$ и $\{d_k\}$.

Задача. *Получить расчетные формулы для* $\{b_k\}$ *и* $\{d_k\}$.

3.3 Сходимость интерполяционных сплайнов

Сформулируем без доказательства [*2)] теорему, устанавливающую характер сходимости *интерполяционного сплайна.* Имеет место

Теорема. *Если* $f(x) \in C^4[a,b]$ *и* $\overline{\omega}_n$ *- равномерная сетка с шагом* $h = \frac{(b-a)}{n}$, *то справедливы оценки:*

$$\begin{aligned} &\|f(x) - s(x)\|_{C[a,b]} = \sup |f(x) - S(x)| \le C_0 M_4 h^4, \\ &\|f'(x) - s'(x)\|_{C[a,b]} \le C_1 M_4 h^3, \\ &\|f''(x) - s''(x)\|_{C[a,b]} \le C_2 M_4 h^2, \end{aligned} \tag{8}$$

где $M_4 = \sup\limits_{[a;b]} |f^{(4)}(x)|$.

[*1)] Об этом далее при рассмотрении решения СЛАУ.
[*2)] см. [3 (Самарский и др.)]

Таким образом, в случае четырежды непрерывно дифференцируемой функции $f(x)$ и краевых условий (s4), имеет место равномерная по x сходимость самого сплайна $s(x)$ и его производных к интерполируемой функции $f(x)$ с указанными порядками точности.

Вывод: *Сплайн-интерполяция* выгодно отличается от *полиномиальной интерполяции сходимостью* и *устойчивостью* вычисления интерполяционного сплайна $s(x)$.

§4. Среднеквадратичная аппроксимация

4.1 Задача аппроксимации функции

Решая задачу о приближении функции на отрезке $[a, b]$ зачастую бывает невыгодно требование интерполяции $g(x_i) = y_i$, т. е. совпадение с приближаемой функцией на некоторой сетке. Особенно это справедливо в случае, когда функция $f(x)$, точнее ее значения в узлах сетки $\{f(x_i)\}$, известны неточно (например, получены в ходе эксперимента).

В таком случае естественна постановка задачи *аппроксимации* на отрезке $[a, b]$, а именно:

Пусть задана функция $f(x)$ и множество функций $\mathcal{F} = \{F(x)\}$ из линейного нормированного (как правило полного) пространства $\mathcal{L}$. Выделяют две естественные проблемы и соответствующие постановки задачи аппроксимации:

а) *аппроксимация с заданной точностью* ε в метрике пространства $\mathcal{L}$: *По заданному* $\varepsilon > 0$ *найти такую функцию* $F_\varepsilon \in \mathcal{F}$ *из* $\mathcal{L}$, *чтобы имело место неравенство:*

$$\|f(x) - F_\varepsilon\|_{\mathcal{L}}^2 < \varepsilon^2;$$

б) Нахождение *наилучшего приближения* на $\mathcal{F}$ (наилучшая аппроксимация в заданной метрике). *Требуется найти* $\overline{F}(x)$, *удовлетворяющую условию*

$$\|f(x) - \overline{F}(x)\|_{\mathcal{L}}^2 = \inf_{\mathcal{F}} \|f(x) - F(x)\|_{\mathcal{L}}^2.$$

(Пространство $\mathcal{F}$ предполагается полным и inf на нём достигается).

Мы ограничимся сначала рассмотрением случая *аппроксимации в гильбертовом пространстве* (*т. е. полном, нормированном, бесконечномерном, евклидовом* пространстве) вещественных измеримых в квадрате с весом $\rho(x) > 0$ функций

$$f(x) \in L_{2,p}[a, b], \quad \text{где} \quad \|f\|^2 = \int_a^b f^2(x)\rho(x)dx < +\infty.$$

В $L_{2,p}[a, b]$ определено скалярное произведение функций f и g:

$$(f, g) = \int_a^b f(x)g(x)\rho(x)dx$$

оно согласовано с соответствующей нормой функции

$$\|f(x)\| = \sqrt{(f, f)}.$$

Соответственно близость функций по $\|\cdot\|_{L_2}$ есть *среднеквадратичная* близость функций на $[a,b]$.
Как известно, в $L_{2,p}[a,b]$ существует ортонормированные системы функций $\{\varphi_k(x)\}$

$$(\varphi_k,\varphi_m)=\delta_{k,m}=\begin{cases}1, & k=m\\ 0, & k\neq m,\end{cases}\qquad k,m\in\mathsf{N}_0.$$

Рассмотрим задачу аппроксимации на линейном многообразии в $\mathcal{L}$, когда в качестве $\mathcal{F}$ рассмотрим линейная оболочка с порождающими элементами $\varphi_0,\ldots,\varphi_n$:

$$\mathcal{F}=\mathrm{Lin}\,(\varphi_0,\ldots,\varphi_n)=\left\{\varphi\in L_{2,p}[a,b];\quad \varphi=\sum_{k=0}^{n}c_k\varphi_k(x);\quad c_k\in R\right\}.$$

В таком случае аппроксимирующая функция $F(x)$ ищется в виде *обобщённого полинома* по системе функций $\{\varphi_k(x)\}$

$$F(x)=\sum_{k=0}^{n}c_k\varphi_k(x).$$

А задача *наилучшего среднеквадратичного приближения* представляет собой задачу приближения *квадратичной функции* $(n+1)$-го переменного $\{c_k\}$

$$\Phi(\bar c_0,\ldots,\bar c_n)=\inf_{\{c_i\}}\left\|f(x)-\sum_{k=0}^{n}c_k\varphi_k(x)\right\|^2_{L_{2,p}[a,b]}.$$

4.2 Существование и единственность наилучшего среднеквадратичного приближения

Вычислим среднеквадратичное уклонение между $f(x)$ и $F(x)$

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

Далее

$$(f,F)=(f,\sum_{k=0}^{n}c_k\varphi_k(x))=\sum_{k=0}^{n}c_k(f,\varphi_k(x))=\sum_{k=0}^{n}c_kf_k;\qquad \text{где}\qquad f_k=(f,\varphi_k)$$

и

$$\|F\|^2=(F,F)=(\sum_{k=0}^{n}c_k\varphi_k(x),\sum_{p=0}^{n}c_p\varphi_p(x))=\sum_k\sum_p c_kc_p(\varphi_k,\varphi_p)=$$
$$\left|\begin{array}{c}(\varphi_k,\varphi_p)=\\ =\delta_{kp}\end{array}\right|=\sum_{k=0}^{n}c_k^2.$$

Тогда

$$\delta^2=\|f-F\|^2=\|f\|^2-2\sum_{k=0}^{n}c_kf_k+\sum_{k=0}^{n}c_k^2+\sum_{k=0}^{n}f_k^2-\sum_{k=0}^{n}f_k^2=$$
$$=\|f\|^2+\sum_{k=0}^{n}(c_k-f_k)^2-\sum_{k=0}^{n}f_k^2;$$

отсюда следует, что

1) наименьшая погрешность на $\mathcal{F}$ достигается, т. е. существует $\overline{F}(x)$;

2) при $\bar{c}_k = f_k$, т. е. на функции

$$\overline{F}(x) = \sum_{k=0}^{n} f_k \varphi_k(x) \equiv F_n(x) \equiv \sum_{k=0}^{n} (f_k, \varphi_k) \cdot \varphi_k(x); \tag{9}$$

погрешность ап+роксимации минимальна;

3) минимальная величина *среднеквадратичной погрешности* равна

$$\delta^2 = \|f - \overline{F}\|^2 = \|f\|^2 - \sum_{k=0}^{n} f_k^2.$$

Таким образом мы показали, что справедливы

1)**Теорема.** *Наилучшее среднеквадратичное приближение обобщенным полиномом по системе функций $\varphi_k(x)$ существует и единственно. Соответствующее приближение дается отрезком обобщенного ряда Фурье по системе $\{\varphi_k(x)\}$.*

2)**Теорема.** *Если система $\{\varphi_k(x)\}$ полна, то построенное приближение $\overline{F}(x) = F_n(x)$ сходится в среднем к $f(x)$ на $[a; b]$.*

Действительно. Из полноты $\{\varphi_k(x)\} \Rightarrow$ равенство Парсеваля - Стеклова

$$\sum_{k=0}^{\infty} f_k^2 = \|f\|^2 = \int_a^b f^2(x)\rho(x)dx,$$

т.е. ряд $\sum f_k^2$ сходится. Откуда

$$\delta^2 = \|f - \overline{F}\|^2 = \|f\|^2 - \sum_{k=0}^{n} f_k^2 = \sum_{k=n+1}^{\infty} f_k^2 \to 0$$

при $n \to \infty$.

Таким образом среднеквадратичное приближение $\overline{F} = F_n(x)$ сходится в среднем к $f(x)$: $F_n \overset{\text{ср}}{\to} f(x)$ и возможна аппроксимация в среднем с любой степенью точности ε

$$\forall \varepsilon > 0 \quad \exists N(\varepsilon) : \forall n > N(\varepsilon) : \|f(x) - F_n(x)\|_{L_{2,p}[a,b]} < \varepsilon.$$

Замечания. Если система функций не ортогональна (но линейно независима — ЛНЗ), то выкладки отчасти усложняются

$$\delta^2 = \Phi(c_0, \ldots, c_n) = \|f(x) - F(x)\|^2 = (f - \sum_{k=0}^{n} c_k\varphi_k, f - \sum_{k=0}^{n} c_k\varphi_k).$$

Из необходимого условия экстремума $\frac{\partial \varphi}{\partial c_i}$ найдем

$$2\left(-\varphi_i(x), f - \sum_{k=0}^{n} c_k\varphi_k\right) = 0 \quad \Leftrightarrow \quad \sum_{k=0}^{n} c_k(\varphi_k, \varphi_i) = (f, \varphi_i). \tag{9*}$$

Мы получим СЛАУ для определения $\{c_k\}$. Ее определитель - определитель Грама линейно независимой системы $\{\varphi_k(x)\}$. Он строго больше нуля

$$G(\varphi_0, \ldots, \varphi_n) = \det \|(\varphi_i, \varphi_k)\| > 0.$$

Это означает, что решения $\overline{F}(x)$ задачи аппроксимации сущетсвует и единственно. Однако заметим, что

1) Особенно плохо то, что увеличение n заставляет измениться все, найденные до этого коэффисиенты $\{c_k\}_{\overline{0,n}}$. В случае ортонормированных (ОНС) или ортогональных (ОС) систем функций $\{\varphi_k(x)\}$ этого не происходит.

2) СЛАУ (9*) с ростом n становится *плохо обусловленной*, ибо $G(\varphi_0, \ldots, \varphi_n) \underset{n\to\infty}{\to} 0$, что приводит к дополнительным рудностям при решении СЛАУ (9*).

3) Численная ортогонализация системы функций $\{\varphi_k(x)\}$ в свою очередь приводит к большой потере точности при проведении процедуры ортогонализации. Поэтому, если n - относительно велико, то нужно стремиться использовать готовые *ортогональные системы* функций. Напомним основные системы ортогональных полиномов на $[a, b]$.

4.3 Ортогональные в L_2 системы полиномов

Напомним основные системы ортогональных полиномов:

1) *многочлены Якоби* $P_n^{(\alpha,\beta)}(x)$ *(степени* n *и порядков* α, β*)* образуют на $[-1, 1]$ ортогональную с весом

$$\rho(x) = (1-x)^\alpha (1+x)^\beta; \qquad \alpha, \beta > -1$$

систему функций.

Они могут быть рекурентно получены по формуле Родрига

$$P_n^{(\alpha,\beta)} = \frac{(-1)^n}{2^n n!}(1-x)^{-\alpha}(1+x)^{-\beta}\frac{d^n}{dx^n}\left((1-x)^{\alpha+n}(1+x)^{\beta+n}\right)$$

$n = 0, 1, \ldots$.

2) *Многочлены Лежандра* $P_n(x)$. *(Частный случай многочленов Якоби:* $\alpha = \beta = 0$, $\rho(x) = 1$*).* Они образуют ортогональную на $[-1, 1]$ с весом $\rho(x) = 1$ систему полиномов

$$P_n(x) = \frac{(-1)^n}{2^n n!}\frac{d^n}{dx^n}\left((1-x^2)^n\right), \quad n = 0, 1, \ldots$$

3) *Многочлены Чебышева* (частный случай многочленов Якоби):

1-го рода: $T_n(x) = \cos(n \arccos x)$, $x \in [-1, 1]$, $n = 0, \ldots, \infty$; $(\alpha = \beta = -\frac{1}{2})$ — образуют ортогональную с весом $\rho(x) = \frac{1}{\sqrt{1-x^2}}$ систему полиномов.

2-го рода: $U_n(x) = \frac{T_{n+1}(x)}{n+1}$, $n = 0, 1, \ldots$ — ортогональная на $[-1, 1]$ с весом $\rho(x) = \sqrt{1-x^2}$ система многочленов.

4) *Многочлены Лагера* $L_n^{(\alpha)}(x)$ *степени* n *и порядка* α образуют на полупрямой $0 \le x < +\infty$ ортогональную с весом $\rho(x) = x^\alpha e^{-x}$; $\alpha > -1$ систему функций

$$L_n^{(\alpha)} = \tfrac{1}{n!}x^{-\alpha}e^x\tfrac{d^n}{dx^n}\left(x^{\alpha+n}e^{-x}\right), \quad n = 0, 1, \ldots$$

5) *Многочлены Эрмита* $H_n(x)$ образуют на прямой $(-\infty < x < \infty)$ ортогональную систему полиномов с весом $\rho(x) = e^{-x^2}$

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n}\left(e^{-x^2}\right)$$

Приведенные системы полиномов могут быть получены ортогонолизацией на $[a, b]$ системы функций $\{x^k\}$ с соответствующим весом.

Помимо рассмотренных ортогональных систем многочленов, часто удобной ортогональной системой функций оказываются решения задачи Штурма-Лиувилля для соответствующего дифференциального эллиптического уравнения 2-го порядка.

Задача. *Записать для рассмотренных полиномов по три первых представителя. Нормировка рассмотренных полиномов.*

§5. Метод наименьших квадратов (МНК)

5.1 Задача среднеквадратичной аппроксимации сеточных функций

Задача среднеквадратичной аппроксимации в случае заданной таблично на сетке $\overline{\omega}_n$ функции приводит к методу, называемому *метод наименьших квадратов*. (Выбор среднеквадратичной аппроксимации всязан с метрикой соответствующего гильбертова пространства сеточных функций).

Рассмотрим сеточный аналог гильбертова пространства $\mathcal{L}_{2,\rho}[a; b]$ — пространство $\mathcal{H}$ сеточных функций на $\overline{\omega}_n$ (это конечномерное "$n+1$"-мерное евклидово пространство) , определив в нем скалярное произведение и норму так:

$$(f, g)_{\mathcal{H}} = \sum_{k=0}^{n} \rho_k f(x_k) g(x_k); \quad \rho_k \geq 0; \qquad \|f\|_{\mathcal{H}} = \sqrt{(f, f)_{\mathcal{H}}}.$$

Мы рассмотрим линейно-независимую систему $\{\varphi_i(x) \,|\, \varphi_i(x) \in \mathcal{L}_{2,\rho}[a; b]\}_N$ [*1)] и будем считать, что с их помощью ищется наилучшее среднеквадратичное приближение обобщенным сеточным полиномом

$$F \equiv F_N = \sum_{i=0}^{N} C_i \varphi_i(x_p); \quad p = \overline{0, n}; \qquad N \neq n, \text{ как правило, } N < n.$$

Такая постановка приводит нас к задаче на экстремум для среднеквадратичного уклонения δ_N^2 на сетке $\overline{\omega}_n$:

$$\Phi \equiv \Phi(\overline{C}_0, \overline{C}_1, \ldots, \overline{C}_N)\delta_N^2 = \|f - F_N\|^2_{\mathcal{L}_{2,\rho}(\overline{\omega}_n)} =$$

$$= \inf_{\{C_i\}} \sum_{k=0}^{n} \rho_k \left(f(x_k) - \sum_{i=0}^{n} C_i \varphi_i(x_k)\right)^2;$$

[*1)] но на самом деле $\mathcal{L}_{2,\rho}[a; b]$ "хорошие" в смысле гладкости функции

Необходимое условие экстремума функции $\Phi(C_0,\dots,C_N)$ — $\frac{\partial\Phi}{\partial C_p}=0 \quad \Leftrightarrow$ дает СЛАУ (т.к. Φ – квадратичная функция) для определения коэффициентов $\{C_i\}_N$. Имеем:

$$\frac{\partial\Phi}{\partial C_p}=\sum_{k=0}^{n}2\rho_k\cdot\left(f(x_k)-\sum_{i=0}^{N}C_i\varphi_i(x_k)\right)\cdot(-\varphi_p(x_k))=0,\ \text{или}$$

$$\Updownarrow$$

$$\sum_{i=0}^{N}C_i(\varphi_i,\varphi_p)_{\mathcal{L}_{2,\rho}(\overline{\omega}_n)}=(f,\varphi_p);\quad p=\overline{0,N} \tag{10}$$

Определитель Грамма $G(\varphi_0,\dots,\varphi_N)\neq 0$ для системы линейно-независимых сеточных функций $\{\varphi_i(x_k)\}_{\substack{i=\overline{0,N}\\k=\overline{0,n}}}$ следовательно решение задачи (10) существует и единственно. (матрица СЛАУ (10) положительная $G>0;\quad (Gx,x)>0;\quad x\neq 0$) Естественно, что трудность решения задачи (10) зависит от системы $\{\varphi_i(x_k)\}$.

Из многочисленных примеров использования метода НК остановимся на его использовании в задаче обработки экспериментальных данных. Рассмотрим сначала

5.2 Обработка экспериментальных кривых методом НК

МНК широко используется в обработке экспериментальных кривых, т.е. таких кривых, точки которых измерены с *известной* погрешностью ε_k — $\{x_k;\ f(x_k)=y_k;\ \varepsilon_k\}$.

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

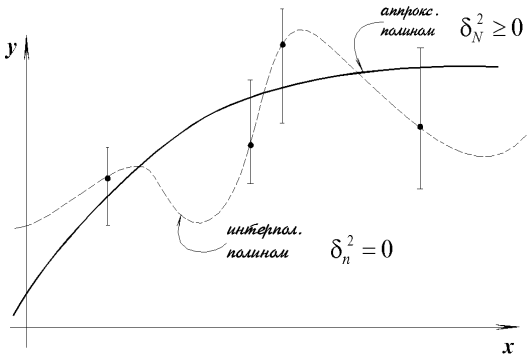
$$\rho_k=\frac{1}{\varepsilon_k^2};$$

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

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

второй сомножитель должен быть существенно меньше для получения той же величины результата.

Естественную интерпретацию получает при этом и проблема выбора числа N членов обобщенного полинома для F_N.



Если число N коэффициентов аппроксимации взять равным числу узлов n сетки $\overline{\omega_n}$, то мы получим задачу интерполяции, как решение задачи об $\inf\delta^2_{N=n}$. Но это при наличии больших экспериментальных ошибок неприемлемо (см. рисунок).

Хорошее "сглаживание" эксперимента будет при $N<n$. Но, если N слишком мало, то коэффициентов (членов ряда для F_N) может не хватить для описания сложной кривой (т.е. δ_N^2 велико).

На практике оптимальное число коэффициентов N определяют следующим образом:

1. Выбирают некоторое N и из системы (10) находят $\left\{\overline{C}_i^{(N)}\right\}_{i=\overline{0,N}}$;

2. Вычисляют получившееся при этом среднеквадратичное уклонение, т.е. величину $\inf \delta_N^2$:
$$\delta_N^2 = \inf_{\{C_i\}} \|f - F_N\|^2 = \Phi\left(\overline{C}_0^{(N)}, ..., \overline{C}_N^{(N)}\right);$$

3. Сравнивают её с известной погрешностью экспериментальных данных $\varepsilon = \max\limits_{k=\overline{0},n} |\varepsilon_k|$

Если $\delta_N \gg \varepsilon$, т.е. "математическая" погрешность много больше "физической" погрешности данных, то число коэффициентов N *недостаточно* для "хорошего" описания $f(x)$ и его нужно *увеличить*: — $N \uparrow$. Если $\delta_N \ll \varepsilon$, то старшие $\left(\begin{smallmatrix}\text{особенно для}\\ \text{ортогональной}\\ \text{системы}\end{smallmatrix}\right)$ коэффициенты ненадежны и нужно уменьшить N — $N \downarrow$.

Оптимально то значение N, при котором $\delta_N \approx \varepsilon$, но $N < n$ (!)
Начинают с $N = 1$. Тогда заведомо $\delta_1 \gg \varepsilon$ и увеличивя N добиваются выполнения условия $\delta_N \approx \varepsilon$, если при этом $N < n$ – то *аппроксимация* достигнута, в противном случае необходим более разумный выбор системы функций $\{\varphi_i(x_k)\}$.

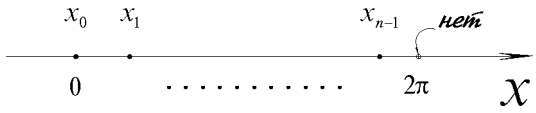
Замечания:

1. В качестве системы сеточных функций $\{\varphi_i(x_k)\}$ часто используют неортогональную систему полиномов $\{(x_k)^i\}$. Соответствующие формулы (10) нсят достаточно простой вид. Однако, вычисления по ним не лишены общего недостатка неортогональных систем $\{\varphi_i(x_k)\}$, связанного с потерей устойчивости нахождения $\{C_i\}_N$ из-за плохой обусловленности СЛАУ (10). Поэтому практически ограничиваются значениями $N \approx 2 \div 5$.

2. Наиболее разумно в методе МНК использовать подходящую ортогональную систему сеточных функций.

Пример. *В качестве примера рассмотрим задачу среднеквадратичной аппроксимации МНК 2π–периодической (или периодического продолжения) функции на равномерной сетке $\overline{\omega}_n^{(h)}$, покрывающей её период $T = [0; 2\pi)$.*

Итак, на отрезке $[0; 2\pi)$ рассматривается сетка $\overline{\omega}_{n-1}$ с узлами $x_p = \frac{2\pi}{n}p$; $p = \overline{0, n-1}$
Покажем, что с весом $\rho_p \equiv 1$ система "тригонометрических" сеточных функций $\varphi_k(x) = exp(ikx)$ ортогональна на $\overline{\omega}_{n-1}^{(h)}$.



Имеем:

$$k \neq m, \quad (\varphi_k, \varphi_m) = \sum_{p=0}^{n-1} \varphi_k(x_p)\overline{\varphi_m(x_p)} \cdot \rho_p^{(\equiv 1)} = \sum_{p=0}^{n-1} \exp\left\{i\frac{2\pi}{n}p(k-m)\right\} =$$

$$= 1 + e^{i\frac{2\pi}{n}(k-m)} + e^{i\frac{2\pi}{n}\cdot 2\cdot(k-m)} + \cdots + e^{i\frac{2\pi}{n}(k-m)\cdot(n-1)} = \left|\begin{smallmatrix}\text{геом. прогрессия с}\\ q = e^{i\frac{2\pi}{n}(k-m)}\end{smallmatrix}\right| =$$

$$= \left|S_n = \frac{a_1 - qa_n}{1-q}\right| = \frac{1 - e^{i\frac{2\pi}{n}(k-m)\cdot n}}{1 - e^{i\frac{2\pi}{n}(k-m)}} = \begin{cases}0, & k \neq m\\ \text{при } k = m\\ (\varphi_m, \varphi_m) = n \text{ (одна и та же}\\ \text{норма у всех функций)}\end{cases}$$

$$\Downarrow$$

$$\underline{(\varphi_k, \varphi_m)_{\overline{\omega}^{(h)}_{n-1}} = n\delta_{k,m}.}$$

В таком случае (10) легко решается

$$\left.\begin{array}{c} (\varphi_k, \varphi_k)C_k = (f, \varphi_k) \;\Leftrightarrow\; C_k = \dfrac{1}{n}\displaystyle\sum_{p=0}^{n-1} f(xp)e^{i(\frac{2\pi}{n}k)\cdot p^{*0)}} \\ \overline{F}(x) = \displaystyle\sum_{k=0}^{N} C_k e^{ikx}. \end{array}\right\} \text{формулы Бесселя} \qquad (11)$$

Благодаря ортогональности систем $\{e^{ikx}\}$ эти функции можно использовать при больших N и n (разумеется нужно стремиться $N \le n-1$). Достаточно просто они выглядят при $n = 12$ (простая тригонометрия,что часто используется).

5.3 Сглаживание (фильтрация) экспериментальных таблиц методом наименьших квадратов

В качестве второй иллюстрации применения МНК в обработке экспериментальных данных рассмотрим задачу сглаживания (или "фильтрации") экспериментальных данных, полученных с *большими ошибками*. При этом поступают так:

Возьмем несколько точек x_i в таблице и в выбранном интервале сетки построим среднеквадратичную аппрксимацию $\varphi(x)$ с одним или двумя параметрами (обычно полиноминальную, т.е. полином невысокой степени).

Центральной точке системы узлов припишем то значение, которое дает аппроксимация.

Для трех точек x_{k-1}, x_k, x_{k+1} при аппроксимации полиномом 1-ой степени (2 параметра) получим(ограничимся случаем *равномерной* сетки $h - const$):

$$\varphi(x) = C_0 + C_1(x - x_k); \qquad x \in [x_{k-1}, x_{x+1}]$$

Погрешность на этих узлах (для всех $\rho_k \equiv 1$):

$$\delta^2 = (\varphi(x_{k-1}) - f(x_{k-1}))^2 + (\varphi(x_k) + f(x_k))^2 + (\varphi(x_{k+1}) - f(x_{k+1}))^2 =$$

$$= (C_0 - hC_1 - y_{k-1})^2 + (c_0 - y_k)^2 + (C_0 + hC_1 - y_{k+1})^2 \equiv \Phi(C_0, C_1)$$

СЛАУ (10) в частном случае для $C_0, C_1 \quad \Leftrightarrow \quad \frac{\partial\Phi}{\partial C_0} = \frac{\partial\Phi}{\partial C_1} = 0$. Имеем:

$$\begin{cases} \cancel{2}(C_0 - \cancel{h}\,C_1 - y_{k-1}) + \cancel{2}(C_0 - y_k) + \cancel{2}(C_0 + \cancel{h}\,C_1 - y_{k+1}) = 0 \\ \cancel{2}(C_0 - hC_1 - y_{k-1}) \cdot (-h) + 0 + \cancel{2}(C_0 + hC_1 - y_{k+1}) \cdot h = 0 \end{cases}$$

Её решение

$$\begin{cases} 3C_0 = y_{k-1} + y_k + y_{k+1} \\ 2h^2 C_1 = h(y_{k+1} - y_{k-1}) \end{cases} \Leftrightarrow \begin{cases} C_0 = \dfrac{y_{k-1} + y_k + y_{k+1}}{3} \\ C_1 = \dfrac{y_{k+1} - y_{k-1}}{2h} \end{cases}$$

$^{*0)} \equiv ik \cdot \frac{2\pi}{n} p \qquad (\text{"}ikx\text{"})$

Окончательно

$$\varphi(x) = \frac{y_{k-1} + y_k + y_{k+1}}{3} + (x - x_k)\frac{y_{k+1} - y_{k-1}}{2h}; \quad k = 1, 2, \ldots, (n-1) \qquad (12)$$

$$\varphi(x_k) = \frac{y_{k-1} + y_k + y_{k+1}}{3}; \quad \begin{array}{l}\text{осреднение по 3-м точкам}\\ \text{(простейший линейный}\\ \text{фильтр)}\end{array} \qquad (*)$$

- В рассмотренном случае сглаживание свелось к *осреднению* табличных значений по соседним узлам;
- В радиотехнике такой способ обработки периодического сигнала называют *фильтрацией*. Можно показать, что формула (*) ослабляет высокочастотную составляющую в спектре сигнала примерно в 3 раза, не изменяя практически его низкочастотной составляющей[*1)].

5.4 О равномерном приближении функций

Заканчивая частичное рассмотрение вопроса среднеквадратичной аппроксимации функций, хотелось бы остановиться на комментарии к равномерному приближении функции. Ищется $\overline{F}$ такая, что

$$\|f - \overline{F}\|_c = \inf_{F \in \mathcal{F}} \sup_{[a,b]} |f(x) - F(x)|$$

Bpdtcnyfz neopema Вейерштрасса решает вопрос о существовании полинома $P_n(x)$ равномерного приближения с заданной точностью непрерывной на $[a, b]$ функции$f(x)$, т.е. $\forall \varepsilon > 0 \quad \exists\, n = n(\varepsilon) > 0$ и $P_n(x)$ такие, что

$$\|f - P_n\|_c = \max_{[a,b]} |f(x) - P_n(x)| < \varepsilon.$$

Мы ограничимся констатацией двух *но*:

1) Пусть для 2π-периодической функции $f(x)$ *тригонометрический* полином наилучшего равномерного приближения есть $P_n(x)$. Доказано, что *тригонометрический* полином среднеквадратичного приближения того же порядка n — $Q_n(x)$ имеет в $C[a, b]$ погрешность не хуже

$$\|f - Q_n(x)\|_{C[a,b]} \le \left(\frac{9}{2} + \ln n\right) \cdot \|f - P_n(x)\|_{C[a,b]}.$$

Тем самым, при разумных n, $Q_n(x)$ ненамного, в смысле приближения, хуже чем $P_n(x)$. Аналогичная по смыслу оценка имеется и для алгебраических полиномов на $[-1, 1]$.

2) Для построения полинома $P_n(x)$ наилучшего равномерного приближения известны *лишь итерационные способы*. В то время как $Q_n(x)$ строится почти явно при решении СЛАУ (10).

[*1)] [3 (Самарский и др.)]

ГЛАВА III

ЧИСЛЕННОЕ ИНТЕГРИРОВАНИЕ

§1. Постановка задачи численного интегрирования

Задача численного интегрирования функций заключается в вычислении приближенного значения определенного интеграла

$$I = \int_a^b f(x)\rho(x)\,dx; \qquad (\rho(x) > 0 \text{ – весовая функция})$$

на основе ряда значений подынтегральной функции $\left\{f(x)|_{x=x_k} = f(x_k) = y_k\right\}$.

Формулы численного вычисления однократного интеграла называются *квадратурными формулами*, двойного и более кратного — *кубатурными*. Ограничимся лишь рассмотрением квадратурных формул.

Обычный и естественный прием построения квадратурных формул состоит в замене подынтегральной функции $f(x)$ на отрезке $[a,b]$ *интерполирующей* или *аппроксимирующей* функцией $g(x)$ сравнительно простого вида (например полиномом) с последующим аналитическим интегрированием. Это приводит к представлению

$$I = \int_a^b f(x)\rho(x)\,dx = \int_a^b g(x)\rho(x)\,dx + R[f], \tag{1}$$

В пренебрежении остаточным членом $R[f]$ получаем приближенную формулу $\tilde{I} = \int_a^b g(x)\rho(x)\,dx$. Мы ограничимся случаем, когда в качестве $g(x)$ выбирается полином (возможно и обобщенный) а $\rho(x) \equiv 1$ (как правило).

Обозначим через $y_i \equiv f(x_i)$ значение подынтегральной функции в различных точках $x_i \in \overline{\omega}_n$ на $[a,b]$[*1)]. В качестве приближенной функции $g(x)$ рассмотрим интерполяционный полином на $\overline{\omega}_n$ в форме полинома Лагранжа:

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

при этом $f(x) = L_n(x) + r_n(x)$; $\Big\|$где $r_n(x)$ – остаточный член интерполяционной формулы Лагранжа ("точная" формула)

Формула (1) дает:

$$\int_a^b f(x)\rho(x)\,dx = \int_a^b L_n(x)\rho(x)\,dx + R_n(f) = \sum_{k=0}^{n} C_k f(x_k) + R_n; \tag{2}$$

[*1)] квадратурные формулы не всегда являются формулами *замкнутого* типа, т.е. $x_0 = a; \quad x_n = b$

где

$$C_k = \int_a^b \frac{\omega(x)}{(x - x_k)\,\omega'(x_k)}\,\rho(x)\,dx, \quad R_n(f) = \int_a^b r_n(x)\rho(x)\,dx. \tag{2'}$$

В формуле (2) величины $\{x_k\}$ – называются узлами, $\{C_k\}$ – весами, R_n – погрешностью квадратурной формулы $\tilde{I} = \sum_{k=0}^{n} C_k f(x_k)$. Если веса C_k вычислены по формуле (2'), то соответствующую квадратурную формулу (2) называют квадратурной формулой *интерполяционного* типа.

Замечания:

1. Веса $\{C_k\}$ квадратурной формулы (2) при заданном расположении узлов $\overline{\omega}_n$ не зависят от вида подынтегральной функции.

2. В квадратурных формулах интерполяционного типа остаточный член $R_n[f]$ может быть представлен в виде значения конкретного дифференциального оператора на функции $f(x)$. Так для $f(x) \in C^{(n+1)}[a,b]$

$$R_n[f] = \int_a^b \frac{f^{(n+1)}(\xi)}{(n+1)!}\,\omega(x)\rho(x)\,dx = C(n^{*1})\frac{f^{(n+1)}(\xi)}{(n+1)!}; \quad \xi \in (a;b).$$

Это позволяет утверждать, что *для полиномов до порядка n включительно квадратурная формула (2) точна, т.е.* $R_n(f) \equiv 0$. Наивысшая степень полинома, для которого квадратурная формула точна, называется *степенью* квадратурной формулы.

Квадратурная формула (2) интерполяционного типа имеет степень не ниже n.

3. Вычисление весов $\{C_k\}$ квадратурной формулы (2) можно проводить не по явным формулам (2'), строя интерполяционный базис Лагранжа, а, используя тот факт, что формула (2) точна для $y = x^k; \quad 0 \le k \le n$. Получим

$$\left.\begin{aligned} I_0 &= \int_a^b x^0 dx = \sum_{k=0}^{n} C_k x_k^0 \\ I_1 &= \int_a^b x^1 dx = \sum_{k=0}^{n} C_k x_k^1 \\ &\cdots\cdots \\ I_n &= \int_a^b x^n dx = \sum_{k=0}^{n} C_k x_k^n \end{aligned}\right\}$$

СЛАУ для $\{C_k\}$

где

$$I_n = \int_a^b x^n\,dx = \frac{x^{n+1}}{n+1}\Big|_a^b$$

Определитель получившейся СЛАУ — есть определитель Вандермонда системы $\{x^k\}$ на сетке $\overline{\omega}_n$

$$\Delta = W(x_0, \ldots, x_n) \neq 0$$

и можно сравнительно компактно записать ответ задачи.

---

*1) зависят от n через шаг сетки

Дальнейшее свое внимание мы сосредоточим на построении интерполяционных квадратурных формул на сетках с постоянным шагом $h = \text{const}$

$$x_{k+1} - x_k = h_{k+1} = h = \frac{b-a}{h} = \text{const}\,.$$

§2. Квадратурные формулы Ньютона-Котесса $\left(\begin{smallmatrix} h=\text{const} \\ \rho\equiv 1 \end{smallmatrix}\right)$

В такой постановке естественно рассматривать квадратурные формулы замкнутого типа. Итак, пусть

$$a = x_0 < x_1 < \ldots < x_n = b; \quad h = \frac{b-a}{n}; \quad x_k = x_0 + kh; \quad k = \overline{0,n}.$$

Тогда при вычислении весовых коэффициентов (2′) $\{C_k\}$ возможны дальнейшие упрощения. Обозначим $\frac{x-x_0}{h} \equiv q$ (выраженная в сеточных шагах длина $x - x_0$), получим

$$\omega(x) = (x-x_0)(x-x_1)\cdots(x-x_n) = h^{n+1}\left(\frac{x-x_0}{h}\right)\cdot\left(\frac{x-(x_0+h)}{h}\right)\cdots$$

$$\cdots\left(\frac{x-(x+nh)}{h}\right) = h^{n+1}q\cdot(q-1)\cdots(q-n);$$

$$\omega'(x_k) = \overbrace{(x_k-x_0)\ldots(x_k-x_{k-1})}^{k\text{ сомножителей}}\cdot\overbrace{(x_k-x_{k+1})\cdots(x_k-x_n)}^{(n-k)\text{ сомножителей}} =$$

$$= h^n k\cdot(k-1)\cdots 1\cdot(-1)\cdot(-2)\cdots(-(n-k)) = (-1)^{n-k}\cdot h^n k!(n-k)!$$

В таком случае

$$C_k = \int_a^b \frac{\omega(x)\,dx}{(x-x_k)\,\omega'(x_k)} = \frac{(-1)^{n-k}}{k!(n-k)!}\,\frac{\not{h}^{n+1}}{\not{h}^{n+1}}\int_a^b \frac{q(q-1)\cdots(q-n)}{(q-k)}\,dx = \left|\begin{matrix}\frac{x-x_0}{h}=q \\ \frac{dx}{h}=dq\end{matrix}\right| =$$

$$= \frac{(-1)^{n-k}}{k!(n-k)!}\,h\int_0^n \frac{q(q-1)\cdots(q-n)}{(q-k)}\,dq; \qquad \text{Окончательно}$$

$$C_k = \frac{(-1)^{n-k}}{k!(n-k)!}\,h\int_0^n \frac{q(q-1)\cdots(q-n)}{(q-k)}\,dq; \qquad \begin{matrix}\text{веса квадратурной формулы} \\ \text{Ньютона-Котесса}\end{matrix} \tag{3}$$

Заменим в (3) $h = \frac{b-a}{n}$ и введем обозначение $C_i = (b-a)\mathcal{K}_i$, тогда

$$\mathcal{K}_i = \frac{(-1)^{n-i}}{i!(n-i!)}\,\frac{1}{n}\int_0^n \frac{q(q-1)\cdots(q-n)}{(q-i)}\,dq; \qquad \begin{matrix}\text{коэффициенты} \\ \text{Котесса}\end{matrix} \tag{4}$$

А сама квадратурная формула принимает вид:

$$\int_a^b f(x)\,dx = (b-a)\sum_{i=0}^{n} f(x_i)\mathcal{K}_i + R_n[f], \qquad \begin{matrix}\text{формула}\\ \text{Ньютона-}\\ \text{Котесса}\end{matrix} \tag{5}$$

где $h = \frac{b-a}{n}$; $\quad f(x_i) = f(a+ih)$.

Замечания:[*1)]

Для коэффициентов Котесса имеют место соотношения:

1. $\sum_{i=0}^{n} \mathcal{K}_i = 1 \qquad$ (ибо для $f \equiv 1$; $R_n[f] = 0$)

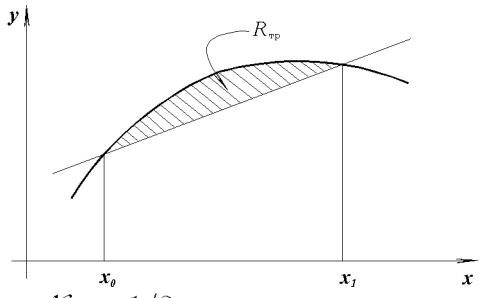
2. $\mathcal{K}_i = \mathcal{K}_{n-i}; \qquad \begin{cases}\text{по построению} + \\ \text{замена} \qquad\quad \text{переменных} \\ q = -\alpha + n\end{cases}$

3. Важно отметить, что коэффициенты Котесса $\mathcal{K}_i$ не являются знакоопределенными, что существенно отразится на свойствах формулы (5) в плане устойчивости суммирования при $n \to \infty$.

§3. Важные частные случаи $n = 1,\ n = 2$

3.1 Квадратурная формула трапеций ($n = 1$)

Пусть соответствующий интерполяционный полином Лагранжа $L_n(x)$ — полином $1^{\text{й}}$ степени.



Тогда

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

что дает квадратурную формулу трапеции

$$\int_{x_0}^{x_1} f(x)\,dx = (x_1 - x_0)\cdot\left(\frac{1}{2}\,y_0 + \frac{1}{2}\,y_1\right) + R_{\text{тр}}; \tag{6}$$

Оценим остаточный член формулы трапеции $R_{\text{тр}}$ как функцию h, в предположении достаточной гладкости $f(x)$. (Это общая техника оценки остаточного члена для

[*1)] 1. и 2. самостоятельно доказать!

интерполяционных квадратурных формул):
Пусть $f(x) \in C^{(2)}[a,b]$. Тогда

$$R \equiv R(h) = \int\limits_{x_0}^{x_0+h} y\,dx - \frac{h}{2}(y_0 + y(x_0+h)), \qquad R(0) = 0. \tag{$*$}$$

Дифференцируя ($*$) по h, найдем

$$R'(h) = y(x_0+h) - \frac{1}{2}(y_0 + y(x_0+h)) - \frac{h}{2}\,y'(x_0+h), \qquad R'(0) = 0.$$

Анологично

$$R''(h) = y'(x_0+h) - \frac{1}{2}\,y'(x_0+h) - \frac{1}{2}\,y'(x_0+h) - \frac{h}{2}\,y''(x_0+h); \qquad R''(0) = 0.$$

Теперь, интегрируя полученное уравнение с соответствующими начальными условиями, найдем

$$R'(h) - R'(0) = \int\limits_0^h R''(t)\,dt = -\frac{1}{2}\int\limits_0^h ty''(x_0+t)\,dt = \left|\begin{smallmatrix}\text{теорема о}\\ \text{среднем}\end{smallmatrix}\right| =$$

$$= -\frac{1}{2}\,y''(\xi)\int\limits_0^h t\,dt = -\frac{h^2}{4}\,y''(\xi); \qquad \xi \in (x_0, h)$$

И окончательно

$$R(h) - R(0) = \int\limits_0^h R'(t)\,dt = -\frac{y''(\xi)}{4}\int\limits_0^h t^2 dt = -\frac{y''(\xi)}{12}\,h^3; \qquad \xi \in (x_0, x_1).$$

Итак,

$$R_{\text{тр}}(h) = -\frac{h^3}{12}\,y''(\xi); \qquad \xi \in (x_o, x_1) \tag{7}$$

Таким образом, мы видим, что (6) формула $1^{\underline{\text{й}}}$ степени, с остаточным членом порядка $O(h^3)$.

3.2 Квадратурная формула Симпсона (формула парабол) ($n = 2$)

Рассмотрим (5) для случая $n = 2$. Сетка $\overline{\omega}_2 = \{x_0, x_1, x_2\}$ содержит три узла. У нас три коэффициента Котесса

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

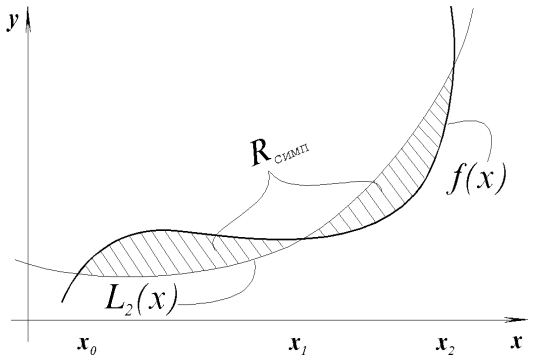
$$\mathcal{K}_0 = \frac{1}{2}\,\frac{(-1)^{2-0}}{0!\,(2-0)!}\int\limits_0^2 \frac{\not{q}(q-1)(q-2)}{\not{q}}\,dq = \frac{1}{4}\int\limits_0^2 (q^2 - 3q + 2)\,dq = \frac{1}{4}\left(\frac{q^3}{3} - \frac{3q^2}{2} + 2q\right)\Bigg|_0^2 = \frac{1}{6}$$

$$\mathcal{K}_0 = \mathcal{K}_2 = 1/6\,; \qquad \mathcal{K}_1 = 1 - 2/6 = 4/6$$

$$\int_{x_0}^{x_2} f(x)\,dx = \overbrace{(x_2 - x_0)}^{2h}\left\{\frac{1}{6}\,y_0 + \frac{4}{6}\,y_1 + \frac{1}{6}\,y_2\right\} + R_{\text{симп.}} = \frac{h}{3}(y_0 + 4y_1 + y_2) + R_{\text{симп.}} \quad (8)$$

Оценка погрешности формулы Симпсона (8) как функции h строится анологично и при $f \in C^{(4)}[a,b]$ можно получить, что

(9) $R_{\text{симп.}} = -\dfrac{h^5}{90} y^{(4)}(\xi)$; *0) $\xi \in (x_0, x_2)$;



Замечания:

1. Формула Симпсона имеет повышенную *степень* при данном числе узлов. Согласно (9) она точна для многочлена до $3^{\text{го}}$ порядка включительно.
2. Случай $n = 3$ дает квадратурную формулу Ньютона. Получить её вид.
3. В общем случае ошибка квадратурной формулы (5) на равномерной сетке для достаточно гладких функций есть

$$R_n[f] = O\left(h^{2[\frac{n}{2}]+3}\right) \quad (10)$$

для формулы с $(n+1)$ узлом интерполяции. Таким образом выгодны формулы с нечетным числом узлов n на сетке.

3.3 Составные квадратурные формулы

Поскольку коэффициенты Котесса K_i громоздки для достаточно больших n, а сам интеграл (1) I обладает свойством аддитивности, то, с практической точки зрения, выгодно использовать составные формулы приближенного интегрирования, разбив отрезок $[a,b]$ на N частичных отрезков и применив на каждом из них квадратурную формулу Ньютона-Котесса (5) невысокого порядка.

а) Общая формула трапеций:

$$a = x_0 < x_1 < \ldots < x_N = b; \quad h = \frac{b-a}{N}.$$

Тогда

$$I = \int_a^b f(x)\,dx = \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} f(t)\,dt = \frac{h}{2}(y_0 + y_1) + \frac{h}{2}(y_1 + y_2) + \ldots + \frac{h}{2}(y_{N-1} + y_N) + R_1 + \ldots + R_N =$$

$$= \frac{b-a}{N}\left(\frac{1}{2}y_0 + [y_1 + \ldots + y_{N-1}] + \frac{1}{2}y_N\right) + R_{tr}. \quad (11)$$

Оценка (7) остаточного члена квадратурной формулы интерполяционного типа, позволяет получить и общую оценку :

$$R_{tr} = \sum_{i=1}^{N} R_i = -\frac{h^3}{12} \sum_{i=1}^{N} y''(\xi_i); \quad \xi_i \in (x_{i-1}, x_i).$$

Рассмотрим среднее арифметическое

$$\mu = \frac{1}{N} \sum_{i=1}^{N} y''(\xi_i).$$

Очевидно, что

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

В силу непрерывности $y''(x)$ на $[a,b]$ теорема Вейерштрасса позволяет утверждать, что

$$\exists\, \xi \in [a,b] : \mu = y''(\xi).$$

Итак :

$$R_{tr} = -\frac{h^3}{12} N \cdot y''(\xi) = -\frac{h^2}{12} (b-a)\, y''(\xi). \tag{12}$$

б) Составная формула Симпсона: Пусть $N = 2m; \quad i = \overline{0, 2m}$, т.е. на отрезке интегрирования находится (2m+1) узел. Применим формулу Симпсона по каждому частичному сдвоенному промежутку :

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

Тогда

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

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

$$+ y_{2m-1}) + 2(y_2 + y_4 + \ldots + y_{2m-2})) + R_{sim}. \tag{13}$$

Если $y \in C^{(4)}[a,b]$, то для оценки остаточного члена формулы (13) получаем :

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

Из аналогичных соображений, использованных при получении формулы (12), найдем, что $\exists\, \xi \in [a,b]$ и

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

что дает представление

$$R_{sim}(h) = -\frac{h^5}{90} \cdot m \cdot y^{(4)}(\xi) = -\frac{h^4}{180}(b-a)y^{(4)}(\xi). \tag{14}$$

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

Замечания: Оценки (12) и (14) позволяют правильно подобрать шаг квадратурной формулы для достижения заданной точности ε при вычислении интеграла в случае достаточно гладких функций, когда известна оценка на $[a,b]$ соответствующей производной. Например, имея оценку для максимума модуля четвертой производной $f(x)$ в области интегрирования — $M_4 = \max_{[a,b]} \left|y^{(4)}(x)\right|$, получим

$$|R_{sim}| \le \frac{b-a}{180} \cdot h^4 \cdot M_4 < \varepsilon \leftrightarrow h \le \sqrt[4]{\frac{180\varepsilon}{(b-a)\cdot M_4}} = O\left(\sqrt[4]{\varepsilon}\right). \tag{15}$$

§4.Апостериорная оценка погрешности квадратурной формулы. Метод Рунге. Метод Эйткена

Во многих случаях вычислений с использованием равномерных сеток (необязательно в задачах вычисления квадратур) погрешности полученных расчетных формул могут быть представлены в виде разложения по степеням шага сетки h. В таком случае можно воспользоваться приемом *повышения точности результата при расчете по формулам фиксированной точности на различных сетках.* Поясним сказанное.

4.1 Метод Рунге (апостериорной оценки точности расчетных формул)

Пусть для вычисления величины $z(x)$ имеется расчетная формула формула $\xi(x,h)$, использующая равномерную сетку с шагом h. И пусть погрешность (остаточный член) этой формулы имеет следующую структуру

$$z(x) - \xi(x,h) = \varphi(x) \cdot h^p + O\left(h^{p+1}\right), \tag{$*$}$$

т.е. погрешность формулы имеет известный порядок p, хотя главный член $\varphi(x)$ асимптотики нам неизвестен. Проведем расчет по той же формуле $\xi(x,h)$, но используя равномерную сетку с другим шагом qh. Получим (с тем же порядком точности) :

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

Имея теперь два расчета $\xi(x,h)$ и $\xi(x,qh)$, нетрудно оценить величину главного члена асимптотического разложения погрешности (*):

$$\varphi(x)\, h^p = \frac{\xi(x,h) - \xi(x,qh)}{q^p - 1} + O\left(h^{p+1}\right), \quad \text{1-я формула Рунге.}$$

Мы видим, что расчет тогоже порядка точности на второй сетке позволяет апостериори оценить погрешность расчета на первой сетке с точностью до членов более высокого порядка по h, т.е. учесть неизвестный порядок $O(h^p)$

$$z(x) = \xi(x,h) + \frac{\xi(x,h) - \xi(x,qh)}{q^p - 1} + O\left(h^{p+1}\right), \quad \text{2-я формула Рунге.}$$

Обычно берут $q = 2$ (что означает сгущение сетки вдвое). Проиллюстрируем применение полученных формул на примере формулы Симпсона (13). Пусть $y^{(4)}(x)$ медленно меняющаяся функция на [a,b]. Тогда можно записать :

$$R_{sim} = -\frac{f^{(4)}(\xi)}{180}(b-a)h^4 \approx Ah^4.$$

Если использовать две сетки с шагом $h_1 = h$ и $h_2 = 2h$, то

$$I \approx I_h + Ah^4 = I_{2h} + A(2h)^4 + O\left(h^5\right) \Longleftrightarrow$$

откуда

$$Ah^4 = \frac{I_h - I_{2h}}{2^4 - 1} = \frac{I_1 - I_{2h}}{15}$$

с тойже точностью $O\left(h^5\right)$. Окончательно

$$\widetilde{I} = I_h + \frac{I_h - I_{2h}}{15} + O\left(h^5\right).$$

4.2 Метод Эйткена (повышения апостериорной оценки точности расчетных формул)

Как способ повышения порядка точности расчетных формул в случае, когда порядок остаточного члена существует, но априори не известен, используют следующий приём.

Рассмотрим, для простоты, три равномерные сетки с шагами $h_1 = h$; $h_2 = qh$; $h_3 = q^2h$. Пусть квадратурная формула такова, что при расчете на сетке h имеем

$$I = I_h + \alpha h^p + O\left(h^{p+1}\right), \qquad (*)$$

т.е. существует асимптотическое разложение остаточного члена $R(h)$ квадратурной формулы.

С точностью до членов более высокого чем $O(h^p)$ порядка по h из отношений (*) можно найти I, α и p:

$$\begin{array}{lcll} I - I_1 & = & \alpha h^p + O\left(h^{p+1}\right) & = A \\ I - I_2 & = & \alpha q^p h^p + O\left(h^{p+1}\right) & = AB \\ I - I_3 & = & \alpha q^{2p} h^p + O\left(h^{p+1}\right) & = AB^2. \end{array}$$

Тогда

$$(I - I_1)(I - I_3) = AB^2 = (I - I_2)^2 \Longleftrightarrow$$

$$I^2 - I(I_1 + I_3) + I_1 I_3 = I^2 - 2II_2 + I_2^2.$$

Окончательно

$$I = \frac{I_2^2 - I_1 I_3}{2I_2 - (I_1 + I_3)} + O(h^{p+1}) \quad \text{— 2-я формула Эйткена.}$$

Оценим эффективный порядок p точности формулы (*), для этого рассмотрим отношение разностей уравнений (*)

$$\frac{I_2 - I_1}{I_3 - I_2} = \frac{A(1-B)}{AB(1-B)} = \frac{1}{B} = \frac{1}{q^p}.$$

Найдем

$$p = \frac{\ln \frac{I_3 - I_2}{I_2 - I_1}}{\ln q} \quad \text{— 3-я формула Эйткена.}$$

Возможно построение и более сложных формул повышения точности выполненных расчётов[*1)].

§5. Квадратурные формулы Гаусса- Кристоффеля

На построение квадратурных формул интерполяционного типа (2') можно посмотреть несколько иначе

$$I = \int_a^b f(x)\, p(x)\, dx = \sum_{i=1}^{n} C_i \cdot f(x_i) + R_n(f) \qquad (2')$$

(здесь удобно суммировать именно с $i = 1$ до n).
Будем считать параметрами квадратурной формулы (2') узлы $\{x_i\}$ и веса $\{C_i\}$ — в нашем распоряжении всего $2n$ парметров. Поставим вопрос *о таком выборе параметров* $\{x_i\}$, $\{C_i\}$, *при которых квадратурная формула (2') точна для многочленов максимально возможного порядка, по крайней мере до* $(2n-1)$ *включительно*[*2)].
Покажем как это сделать.

5.1 Выбор узлов квадратурной формулы $\{x_i\}$

Будем считать что вес $\rho(x)$ непрерывен на $[a,b]$. Он может обратиться в ноль или $+\infty$ лишь в граничных точках отрезка. Известно, что

1) Для такого веса $\rho(x)$ существует полная в $L_{2,p}[a,b]$ система алгебраических полиномов $\{P_k(x)\}_{k=\overline{0,\infty}}$, ортоганальная на $[a,b]$ с весом $\rho(x)$, т.е.

$$\int_a^b P_k(x)\, P_m(x)\, \rho(x)\, dx = \delta_{k,m} \|P_k(x)\|^2_{L_{2p}[a,b]};$$

2) Все нули многочлена $P_k(\mu) = 0 \Longleftrightarrow \left\{\mu_i^{(k)}\right\}_{i=\overline{1,k}}$ — действительны и расположены на интервале (a,b).

[*1)] см.[2 (Бахвалов и др.)] *метод Ричардсона*
[*2)] т.е. имеет максимальную *степень*

Поступим следующим образом :
Составим по неизвестным пока узлам интегрирования $\{x_i\}_{i=\overline{1,n}}$ многочлен n-ой степени

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

Тогда функция $\varphi(x) = \omega(x) P_m(x)$ при $m \leq n-1$ есть многочлен степени k не выше, чем $2n-1$. Для таких многочленов формула (2′) Гаусса-Кристоффеля точна (по предположению):

$$\int_a^b \omega(x) P_m(x) \rho(x) dx = \sum_{k=1}^{n} C_k \omega(x_k) P_m(x_k) \equiv 0 \Leftrightarrow \omega(x) \perp P_m(x), \qquad (*)$$

т.к. $\omega(x_k) = 0, \forall m \leq n-1$ в силу своего построения.

Следовательно многочлен $\omega(x)$ ортогонален линейной оболочке из функций $P_m(x)$ с $m \leq n-1$. Тем самым $\omega(x)$ ортогонален любому многочлену степени $m \leq n-1$.

С другой стороны, если разложить $\omega(x)$ в ряд по ортогональным многочленам $\{P_k(x)\}$

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

то (*) можно продолжить

$$(\omega(x), P_m(x)) = 0 = \sum_{k=0}^{n} a_k \int_a^b P_k(x) P_m(x) \rho(x) dx = \sum_{k=0}^{n} a_k \delta_{k,m} \|P_k\|^2 = a_m \|P_m\|^2.$$

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

$$\omega(x) = A \cdot P_n(x).$$

Мы получили возможность сформировать важный вывод :

1) Узлы $\{x_i\}$ квадратурной формулы Гаусса-Кристоффеля нужно выбирать так, чтобы они совпадали с корнями ортогонального на [a,b] с весом $\rho(x)$ многочлена $P_n(x)$:

$$P_n(\mu) = 0 \Leftrightarrow x_i = \mu_i^{(n)}; \quad i = \overline{1,n}. \qquad (15)$$

5.2 Веса $\{C_i\}$ квадратурной формулы Гаусса-Кристоффеля

Веса квадратурной формулы (2') нетрудно определить, если узлы $\{x_i\}$ уже известны. Для функций $l_m(x)$ — базиса интерполяционных полиномов Лагранжа

$$l_m(x) = \frac{\omega(x)}{(x - x_m)\omega'(x_m)},$$

полиномов $(n-1)$ степени формула Гаусса-Кристоффеля (Г-К) (2′) точна. Тогда

$$\int_a^b l_m(x) \rho(x) dx = \sum_{i=1}^{n} C_i \underbrace{l_m(x_i)}_{\delta_{i,m}} = C_m. \qquad (16)$$

Полученная формула (16) позволяет утверждать, что (2′) есть квадратурная формула *интерполяционного типа.*

Замечания: Весовые коэффициенты $\{C_i\}$ (16) формулы Гаусса-Кристоффеля обладают рядом интересных и важных свойств :

1) Рассмотрим функцию $l_m^2(x) \geq 0$. Она многочлен $(2n-2)$-ой степени $\Rightarrow$ формула (2') точна и поэтому :

$$\int_a^b l_m^2(x)\,\rho(x)\,dx = \sum_{k=1}^{n} C_k l_m^2(x_k) = C_m > 0, \quad \forall m.$$

2) Если интегрировать $f(x) \equiv 1$, то

$$\int_a^b f(x)\,\rho(x)\,dx = \sum_{k=1}^{n} C_k = M, \quad \forall n.$$

Совокупность коэффициентов $\{C_k(n)\}$ равномерно по n ограничена.

3) Формулы Гаусса-Кристоффеля называются формулами наивысшей алгебраической степени, поскольку для *произвольного* многочлена степени большей, чем $(2n-1)$, формула с n узлами не может быть точна.

4) *(О погрешности формулы Гаусса-Кристоффеля).* Погрешность формулы (2′) Гаусса-Кристоффеля пропорциональна производной порядка низшей неучтенной степени интерполяционного многочлена. Для верхней границы погрешности имеем оценку:

$$|R_n| \approx \left(\frac{2}{5}\right)\frac{b-a}{\sqrt{n}}\left(\frac{b-a}{3^n}\right)^{2n} \cdot M_{2n}; \qquad \text{где} \qquad M_{2n} = \max_{[a,b]} \left|f^{(2n)}(x)\right|.$$

Формула Гаусса-Кристоффеля расчитана на интегрирование достаточно гладких функций.

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

Известно, что весу $\rho(x) \equiv 1$ на $[-1,1]$ отвечает система ортогональных полиномов Лежандра $L_n(x)$. В таком случае

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

$$= c\int_{-1}^{1} F(t)dt \approx c \cdot \sum_{k=1}^{n} \gamma_k F(t_k); \quad F(t_k) \equiv f(x_k).$$

Зная узлы t_k и веса γ_k находим

$$x_k = \frac{b-a}{2}t_k + \frac{b+a}{2}$$

и

$$c_k = \frac{b-a}{2} \cdot \gamma_k, \quad k = \overline{1,n}$$

узлы и веса исходной квадратурной формулы.

Рассмотрим простейший возможный случай одного узла $n = 1$. Соответствующий многочлен Лежандра $L_1(t) = t$ (проверить !). Его корень: $t = 0 \leftrightarrow t_1 = 0, \mu_1 = 0$, т.о. $x_1 = (b+a)/2$. Вес этого слагаемого в интегральной сумме:

$$\gamma_1 = \int_{-1}^{1} l_1(t)\,dt = \int_{-1}^{1} \frac{(t-t_1)}{(t-t_1)} dt = 2; \quad \Longrightarrow C_1 = \frac{b-a}{2}\gamma_1 = b-a.$$

Итак

$$\int_a^b f(x)\,dx = (b-a)\, f\left(\frac{a+b}{2}\right) + R_1 \tag{17}$$

формула средних прямоугольников. (Формула открытого типа, точна для многочленов до 1-го порядка включительно).

Задание. *Получить составную формулу средних прямоугольников — формулу (18). Получить оценку погрешности формулы (17).*

§6. Корректность задачи численного интегрирования

Корректность задачи численного интегрирования (2) связана с устойчивостью вычисления интегральной суммы :

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

При этом погрешность $\delta f(x)$ подынтегральной функции обуславливает погрешность δI_n интегральной суммы I_n.

$$I_n + \delta I_n = \sum_{k=0}^{n} C_k \left(f(x_k) + \delta f(x_k)\right) \Leftrightarrow \delta I_n = \sum_{k=0}^{n} C_k \delta f(x_k).$$

Поскольку все рассмотренные нами квадратурные формулы были для $f = 1$ точны, то

$$\int_a^b \rho(x)\,dx = M = \sum_{k=0}^{n} C_k > 0,$$

и имеет место равномерная по n ограниченность $\sum^{n} C_k$ последовательности частичных сумм. Тогда

$$|\delta I_n| \le \sum_{k=0}^{n} |C_k|\,|\delta f(x_n)| \le \max_{[a,b]} |\delta f(x_k)| \cdot \sum_{k=0}^{n} |C_k|.$$

Теперь, если все весовые коэффициенты квадратурной формулы (2) знакопостоянны (в частности $C_k > 0$ для формул Г-К), то можно продолжить :

$$|\delta I_n| \leq \max_{[a,b]} |\delta f(x_k)| \cdot M = M \cdot \|\delta f\|_C .$$

В этом случае δI_n имеет тот же порядок, что и погрешность в вычислении функции, т.е. вычисления по квадратурной формуле устойчивы.

Если же C_k не знакопостоянны, то может оказаться, что абсолютной сходимости ряда $\sum C_k$ нет и ряд $\sum |C_k|$ расходится.(Нет равномерной по n ограниченности у величины $\sum^{\infty} |C_k|$ тем самым δI_n может с ростом n неограниченно возрастать).

Отсутствием знакоопределенности обладают коэффициенты Котесса K_i в формуле (4). Тем самым требуется известная аккуратность при использовании формул Ньютона-Котесса при больших n.

§7. Особые случаи использования квадратурных формул

Особые случаи использования квадратурных формул охватывают важные для практических приложений ситуации применения квадратурных формул (перечислим некоторые из них) :

1) подынтегральная функция не обладает достаточной гладкостью, например кусочно-непрерывная;

2) для подынтегральной функции линейная интерполяция или аппроксимация полиномами неточна. Возможно необходимо рассматривать другие методы интерполяции;

3) вычисляется несобственный интеграл;

4) вычисляется интеграл с переменным верхним пределом;

и т.д.

Остановимся на:

7.1 Интегрирование быстроосциллирующих функций (метод Филона)

Рассмотрим интеграл вида :

$$I = \int_a^b f(x) e^{iwx} dx \equiv \int_a^b F(x) dx,$$

при этом область интегрирования такова, что $w(b-a) \gg 1$; $f(x)$ - достаточно гладкая функция; $w = \text{const}$ — "большая" величина.

Функции $Re\left(f(x) e^{iwx}\right)$, и $Im\left(f(x) e^{iwx}\right)$ имеют на $[a,b]$ примерно $\frac{w(b-a)}{\pi}$ нулей, ибо фазовый множитель изменяется в области интегрирования на величину порядка $w(b-a)$. Число "периодов" $F(x)$ $\sim$ $\frac{w(b-a)}{2\pi}$; на каждом периоде — 2 корня. Таким образом в области интегрирования порядка $\frac{w(b-a)}{\pi}$ - нулей.

Производная функции $F(x)$ при такой постановке, в главном порядке по w есть величина

$$F^{(p)}(x) \sim w^p$$

и для хорошей аппроксимации интеграла и подынтегральной функции приходится выбирать многочлен высокой степени (в любом случае много узлов сетки).

Из оценки остаточного члена для квадратурных формул интерполяционного типа (например на равномерных сетках (12) и (14)) имеем

$$(wh) \ll 1, \text{ т.е. } h \ll \frac{1}{w}.$$

Это означает, что величина шага $h = \frac{b-a}{n} \ll \frac{1}{w}$ должна быть мала или, что то же самое, должно быть велико $n \gg w\,(b-a) > \frac{w(b-a)}{\pi}$. Таким образом на каждом периоде функции $F(x)$ необходимо брать много узлов сетки. Необходима густая сетка и высокой степени полином, что невыгодно с любой точки зрения – большой объем вычислений, громоздкие формулы и т.д.

Естественно желание построить разумные составные квадратурные формулы. Мы воспользуемся некоторой априорной информацией о поведении амплитуды $f(x)$ подынтегральной функции. Если амплитуда $f(x)$ медленно меняется за период фазового сомножителя, то можно использовать составные квадратурные формулы, в которых на каждом частичном интервале (порядка периода и больше !) используется интерполяционный многочлен для $f(x)$ невысокой степени, и дальнейшее интегрирование выполняется *точно!* Такой подход приводит к *квадратурным формулам Филона.*

Пусть интервал $[a, b]$ разбит на N частей точками

$$a = x_0 < x_1 < ... < x_N = b.$$

На k-ом частичном интервале $[x_{k-1}, x_k]$ построим интерполяционный полином $P_q\,(x)$ по $(q+1)$ узлу (необязательно замкнутого типа) для $f(x)$. Тогда :

$$I_N = \sum_{k=1}^{N} I_k = \sum_{k=1}^{N} \int_{x_{k-1}}^{x_k} P_q\,(x)\, e^{iwx} dx.$$

Собственно формулы Филона получаются при $q = 2$ (при квадратичной интерполяции). Мы ограничемся случаем $q = 1$ (линейная интерполяция). Запишем интерполяционный полином в форме интерполяционного полинома Ньютона на тех же узлах $\{x_{k-1}, x_k\}$.

$$P_1\,(x) = N_1\,(x) = f\,(x_{k-1}) + (x - x_{k-1})\, f\,(x_{k-1}, x_k) = y_{k-1} + \frac{y_k - y_{k-1}}{x_k - x_{k-1}}\,(x - x_{k-1})\,.$$

Тогда

$$I_k = \int_{x_{k-1}}^{x_k} N_1\,(x)\, e^{iwx} dx = \int_{x_{k-1}}^{x_k} \left(y_{k-1} + \frac{y_k - y_{k-1}}{x_k - x_{k-1}}\,(x - x_{k-1}) \right) e^{iwx} dx =$$

$$= y_{k-1} \frac{e^{iwx}}{iw} \Big|_{x_{k-1}}^{x_k} + \frac{y_k - y_{k-1}}{x_k - x_{k-1}} \left\{ (x - x_{k-1}) \frac{e^{iwx}}{iw} \Big|_{x_{k-1}}^{x_k} - \int_{x_{k-1}}^{x_k} \frac{e^{iwx}}{iw} dx \right\} =$$

$$= \frac{y_{k-1}}{iw} \left(e^{iwx_k} - e^{iwx_{k-1}} \right) + (y_k - y_{k-1}) \frac{e^{iwx_k}}{iw} - \frac{y_k - y_{k-1}}{x_k - x_{k-1}} \frac{e^{iwx}}{(iw)^2} \Big|_{x_{k-1}}^{x_k} =$$

$$= -\frac{F_{k-1}}{iw} + \frac{F_k}{iw} + \frac{y_k - y_{k-1}}{w^2 h_k}\left(e^{iwx_k} - e^{iwx_{k-1}}\right) = \left| \begin{array}{c} x_{k-1} = x_k - \frac{h_k}{2} - \frac{h_k}{2} = x_{k-1/2} - \frac{h_k}{2} \\ x_k = x_k - \frac{h_k}{2} + \frac{h_k}{2} = x_{k-1/2} + \frac{h_k}{2} \end{array} \right| =$$

$$= \frac{F_k - F_{k-1}}{iw} + \frac{y_k - y_{k-1}}{w^2 h_k} e^{iwx_{k-1/2}} \cdot \sin\frac{wh_k}{2} \cdot 2i;$$

Теперь осталось просуммировать I_k по k:

$$I_N = \sum_{k=1}^{N} I_k = \frac{1}{iw}\left(F_1 - F_0 + F_2 - F_1 + ... + F_N - F_{N-1}\right) + \frac{2i}{w^2}\sum_{k=1}^{N} \frac{\sin\frac{wh_k}{2}}{h_k}\left(y_k - y_{k-1}\right) e^{iwx_{k-1/2}} =$$

$$= \frac{F_N - F_0}{iw} + \frac{2i}{w^2}\sum_{k=1}^{N} \frac{\sin\frac{wh_k}{2}}{h_k}\left(y_k - y_{k-1}\right) e^{iwx_{k-1/2}} \quad \text{- формула Филона.} \qquad (19)$$

Замечания:

а) упростить полученную формулу для случая равномерной сетки $h_k = h = const$ ... $\Rightarrow$ формула (20).

б) при $hw \ll 1$ формула (19), (20) переходит в обобщенную формулу трапеций (что естественно) и имеет погрешность $R = O\left(h^2\right)$. Однако это требует рассмотрения слишком малого шага $h \ll 1/w$.

Если же $\frac{1}{w} < h \ll 1$, то погрешность формулы (19) имеет порядок $R = O\left(\frac{Nf''(\xi)}{w^3}\right)$ [*1)] и она малая величина, если $f''(\xi)$ мало, т.е. характер изменения амплитуды $f(x)$ близок к линейному.

В таком случае возможно интегрирование с достаточно большим шагом $h > \frac{1}{w}$ (порядка длины волны и более).

[*1)] из $R_k = \int_{x_{k-1}}^{x_k} f''(\xi)/(2!)(x - x_{k-1})(x - x_k)e^{iwx}dx \sim \frac{1}{w^3}$ следует оценка остаточного члена в формуле Филона

ГЛАВА IV

ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ НЕЛИНЕЙНЫХ УРАВНЕНИЙ

§1.Постановка задачи. Метод простой итерации

Одной из наиболее распространённых вычислительных задач является задача нахождения *корня уравнения.*
Пусть $f(x)$ - непрерывная и достаточно гладкая функция действительного переменного x. Нас будет интересовать проблема нахождения всех или нескольких корней уравнения

$$f(x) = 0. \tag{1}$$

Естественное обобщение уравнения (1) на случай рассмотрения вектор-функции от m-мерного аргумента $x \in R^m$ приводит к задаче отыскания решения системы нелинейных уравнений:

$$f_i(x_1, ..., x_m) = 0, i = \overline{1, m}. \tag{1'}$$

Мы ограничимся именно такой постановкой $i = \overline{1, m}$ (предполагая некую замкнутость по числу неизвестных и уравнений).

В решении задачи (1) и (1′) выделяют несколько характерных этапов:

1) Нужно исследовать количество; характер (подразумевая под этим кратность) и расположение (локализацию) корней;

2) Выбор начального приближения. Поскольку алгоритм решения задач (1), (1′) представляет собой, как правило, итерационную процедуру построения последовательности $\{x_k\}$, $x_k \to x^*$, сходящейся к корню x^* уравнения (1), то особую роль играет выбор начального приближения x_0;

3) Собственно построение самой последовательности $\{x_k\}$;

Ограничемся случаем отыскания действительных корней, хотя уже в случае алгебраического уравнения (1) $f(x) \equiv p_n(x) = 0$ их может и не быть.

На примере уравнения (1), т.е. в случае уравнения с одним неизвестным, вопрос о локализации корня и об отделении корней решают, как правило, *методом дихотомии*, т.е. последовательным делением отрезка пополам. При этом в качестве нового частичного отрезка, содержащего корень, берут интервал вдвое меньшей длины, на котором происходит смена знака. Пусть смена знака происходит на интервале $[x_l; x_r]$. Тогда

$$f(x_l) \cdot f(x_r) \le 0;\ f\left(x_{cp} = \frac{x_l + x_r}{2}\right) \cdot f\left(x_{(l?r)}\right) \le 0.$$

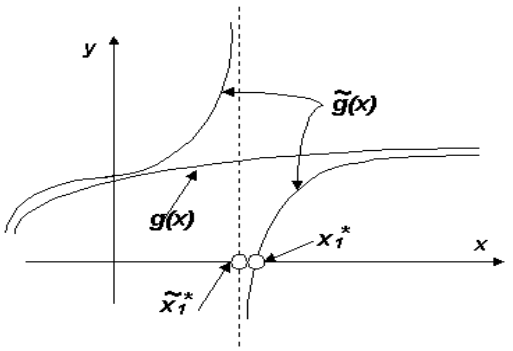
Проведенные таким образом несколько итераций позволяют локализировать часть корней (1) на интервалах некоторой сетки $\{x_i\}$.

После того, как один из корней, скажем x_1^*, найден - *выделяют* данный корень, т.е. переходт к рассмотрению функции $g(x)$:

$$g(x) = \frac{f(x)}{(x - x_1^*)}.$$

Если $f(x) \in Lip^{*1)}([a,b],c)$ — *липшиц-непрерывная* функция, то $g(x)$ является *непрерывной* функцией и корни $g(x)$ совпадают с корнями $f(x) = 0$ за исключением x_1^*, либо x_1^* имеет на единицу меньшую кратность.

На практике выделение корня нужно проводить аккуратно, с высокой точностью, ибо, поскольку мы находим приближенное значение корня $\widetilde{x}_1^*$, то функция $\widetilde{g}(x) = \frac{f(x)}{x - \widetilde{x}_1^*}$ имеет в точке x_1^* ноль и в точке $\widetilde{x}_1^*$ - полюс. Тем самым "сильно" отличается от $g(x)$ в некоторой окрестности x_1^*.



Возможность построения последовательности $x_k \to x^*$ основана, как правило, на рассмотрении задачи (1), (1′) как задачи "о неподвижной точке" некоторого отображения $\varphi(x)$

$$x = \varphi(x). \tag{2}$$

Обычно от (1) $\Rightarrow$ (2) переходят таким образом:

Умножим (1) на *непрерывную, достаточно гладкую, знакопостоянную* в G (области локализации корня) функцию $\tau(x)$ и добавим тождество $x = x$. Получим:

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

Приближенное решение задачи (2) строится, что естественно, методом *простой итерации* (или методом *последовательных приближений* — МПП) :

$$\begin{cases} x_{k+1} = \varphi(x_k), & k = 0, 1, 2, ... \\ x_0 = \overset{(0)}{x} & \text{- дано} \end{cases} \tag{3}$$

— *одношаговый итерационный метод.*

Говорят, что итерационная последовательность (3) $\{x_k\}$ обладает сходимостью p-го порядка к корню x^* (1), если для погрешности следующей итерации имеем оценку:

$$\|x_{k+1} - x^*\| = O\left((x_k - x^*)^p\right); \quad p\text{- может быть и нецелым.}$$

Мы ограничемся рассмотрением итерационных последовательностей, обладающих линейной или квадратичной сходимостью.

*1) $f(x) \in Lip(G;c)$ если $\exists c$; $\forall x_1, x_2 \in G$ $|x_1 - x_2| < \delta$, имеем $|f(x_1) - f(x_2)| \le c|x_1 - x_2|$, т.е. её модуль непрерывности $\varepsilon(\delta) \le c \cdot \delta$.

§2.Сходимость метода простой итерации

Для наглядности рассмотрим обобщение задачи (2) на случай *непрерывного* отображения *полного* метрического просранства R в себя. При этом *отображение* $\varphi : R \Rightarrow R$ *называют сжимающим (сжатием в* R*), если* $\exists\, q \in (0,1)$*:*

$$\forall x, y \in R : \rho(\varphi(x), \varphi(y)) \leq q \cdot \rho(x, y)$$

Построение последовательности $x_k \to x^*$ основано на теореме о неподвижной точке:

Теорема.1 (*принцип сжимающего отображения*) *Если* φ *- непрерывное сжатие полного метрического пространства* R*, то существует единственная неподвижная точка* $x^* : x^* = \varphi(x^*)$ *и она является пределом последовательности:*

$$x_{k+1} = \varphi(x_k), \quad k = 0, 1, \ldots \quad \forall x_0 \in R$$

$$\lim_{k\to\infty} x_k = x^*.$$

Доказательство: (Поскольку оно не связано с устройством конкретного пространства, то оно не столь громоздко и многие полезные моменты становятся нам яснее).

1) Покажем, что последовательность $\{x_k\}$ (3) *фундаментальна* в R. Пусть m, n — произвольные натуральные числа $(m > n)$, тогда

$$\rho(x_m, x_n) = \rho(\varphi(x_{m-1}), \varphi(x_{n-1})) \leq q \cdot \rho(x_{m-1}, x_{n-1}) \leq \ldots \leq q^n \rho(x_{m-n}, x_0).$$

Теперь оценим $\rho(x_{m-n}, x_0)$ через расстояние между соседними точками :

$$\rho(x_{m-n}, x_0) = \rho(x_0, x_{m-n}) \leq \rho(x_0, x_1) + \rho(x_1, x_2) + \ldots + \rho(x_{m-n-1}, x_{m-n}) \leq$$

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

$$\leq \rho(x_0, x_1)\left(1 + q + q^2 + \ldots + q^{m-n-1} + \ldots\right) = \rho(x_0, x_1)\frac{1}{1-q} \text{ — (так далеко ушли).}$$

Тогда

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

и не зависит от m для любого n, т.о. $\{x_n\}$ — *фундаментальна.*

2) В силу *полноты* R у $\{x_n\}$ есть предел, пусть $\lim_{k\to\infty} x_k = x^*$. Покажем, что это неподвижная точка непрерывного отображения φ, действительно:

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

Итак

$$x^* = \varphi(x^*).$$

3) Покажем, что x^* — единственная неподвижная точка отображения φ в R. Допустим противное и назовем эти точки $\overline{x}$ и $\overline{\overline{x}}$. Точки $\overline{x}$ и $\overline{\overline{x}}$ неподвижные точки для отображения φ, тогда:

$$\rho\left(\overline{x}, \overline{\overline{x}}\right) > 0, \text{ но } \rho\left(\overline{x}, \overline{\overline{x}}\right) = \rho\left(\varphi\left(\overline{x}\right), \varphi\left(\overline{\overline{x}}\right)\right) \leq q \cdot \rho\left(\overline{x}, \overline{\overline{x}}\right) \quad \Rightarrow \quad q \geq 1.$$

Что невозможно, ибо φ - сжатие ■

Обычно задача (2) рассматривается нами локально, т.е. в ситауции, когда корень x^* локализован. Вопрос о локализации корня решает

Теорема 2. Пусть область G — открытая область полного метрического пространства в R и пусть на G задано сжимающее отображение в R — φ: $G \Rightarrow R$. Тогда для существования неподвижной точки x^* отображения φ в области G, $x^* \in G$ необходимо и достаточно, чтобы нашелся в области G замкнутый шар $\overline{K}(x_0, r) \subset G$, т.е.

1) нашлось положительное число $r > 0$;

2) и точка $x_0 \in G$ — такая, что замкнутый шар $\overline{K}(x_0, r) \subset G$ и имеет место неравенство

$$\rho(x_0, \varphi(x_0)) \leq (1-q)\, r.$$

(Образ центра оставался бы в шаре, т.е. $\overline{K}(x_0, r)$ - должным образом "центрирован").

Доказательство. (Необходимость)

Пусть в G существует неподвижная точка x^* для отображения φ. Поскольку G открытая область, то x^* - внутренняя точка области $G \Rightarrow$ существует шар $\overline{K}(x^*, r) \subset G$ (здесь $r < dist(x^*, \partial G)$)).

Тогда

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

и подавно для $x_0 = x^*$.

Достаточность. В силу условий теоремы существуют такие x_0 и $r > 0$, что замкнутый шар $\overline{K}(x_0, r) \subset G$. Покажем, что $\varphi(\overline{K}(x_0, r)) \subseteq \overline{K}$, т.е. φ - сжатие внутри $\overline{K}$. Действительно:

$$\forall x \in \overline{K}(x_0, r) : \rho(\varphi(x), x_0) \leq \rho(\varphi(x), \varphi(x_0)) + \rho(\varphi(x_0), x_0) \leq$$

$$\leq q\rho(x, x_0) + (1-q)\, r \leq r, \text{ т.о. } \varphi(x) \in \overline{K}(x_0, r).$$

Поскольку $\overline{K}(x_0, r)$ — замкнутое полное подпространство в R и в нём выполнены условия *Теоремы 1*, то $\Rightarrow$ существует единственная точка $x^* \in \overline{K}(x_0, r)$ такая, что

$$x^* = \varphi(x^*).$$

Подавно $x^* \in G$ ■

Замечания:

1) В условиях Т.2 не нужно рассматривать отображение φ на всем R. Это означает, что мы локализовали x^* - корень уравнения (1) в пределах шара $\overline{K}(x_0, r)$.

2) Метод последовательных приближений обеспечивает не хуже, чем линейную сходимость $\{x_k\}$:

$$\rho(x_{k+1}, x^*) = \rho(\varphi(x_k), \varphi(x^*)) \leq q \cdot \rho(x_k, x^*) \leq \ldots \leq q^{k+1}\rho(x_0, x^*); \tag{4}$$

3) Оценка погрешности на n-ой итерации следует из фундаментальности $\{x_k\}$:

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

Переходя к пределу, получим

$$\lim_{m\to\infty} \rho(x_n, x_m) = \rho(x_n, x^*) \leq \frac{q^n}{1-q}\rho(x_0, x_1); \tag{5}$$

(оценка не содержит неизвестной точки x^*.)

§3. Итерационные методы решения уравнения $f(x) = 0$ с одним неизвестным

Одношаговые итерационные методы удобно записать в виде *метода простой итерации:*

$$\begin{cases} x_{n+1} = \varphi(x_n) = x_n + \tau(x_n) f(x_n); \quad n = 0, 1, 2, \ldots \\ x_0 \text{ - дано.} \end{cases}$$

Среди них

а) *Метод релаксации*: $\tau(x_n) = \tau - const$, т.е. стационарный метод

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

(получается из решени на установление задачи, для дифференциального уравнения $\dot{x} = f(x)$)

б) *Метод Ньютона*: (метод линеаризации или метод касательных)

Получается заменой уравнения (1) $f(x) = 0$ "близким" ему уравнением в окрестности приближения x_n из разложения $f(x)$ в ряд Тейлора до членов 1-го порядка включительно:

$f(x) = f(x_n) + f'(x_n)(x - x_n) + R_1$, что дает приближенное уравнение

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

или

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

носит название метода *касательных*[*1)].

Модифицированный метод Ньютона:

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

Реализуя метод, избегают многократного вычисления $f'(x)$ (только на шагах обновления x_p).

в)*Метод секущих*: Этот метод получают либо из метода Ньютона заменой $f'(x_n)$ разделенной разностью

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

Либо рассматривая *интерполяционный метод* 1-ого порядка :

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

[*1)]уравнение касательной в точке $(x_n, y(x_n))$ имеет вид $y = f(x_n) + f'(x_n)(x - x_n)$

т.е.

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

Получаем алгоритм

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

x_0, x_1 — дано. Мы получили *двухшаговый* итерационный метод.

г) *Метод парабол* - итерполяционный метод 2-го пордка. Пусть приближения x_{n-2}, x_{n-1}, x_n - известны. Строим интерполяционный полином

$$N_2(x) = f(x_n) + (x - x_n) f(x_n, x_{n-1}) + (x - x_n)(x - x_{n-1}) f(x_n, x_{n-1}, x_{n-2}).$$

Из уравнения

$$N_2(x) = 0$$

— это квадратное уравнение относительно $x - x_n = z, \quad az^2 + bz + c = 0$. Определим очередное приближение x_{n+1}. Пусть z_n наименьший по модулю корень[*1)], тогда

$$x_{n+1} = x_n + z_n.$$

Метод парабол позволяет отыскивать *комплексные корни* при действительном начальном приближении.

д) *Метод обратной интерполяции*: Этот итерационный метод получают с помощью интерполяции обратной к $f(x)$ функции $x = g(y)$

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

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

$$y_0 = f(x_0); \ldots; y_n = f(x_n).$$

Построим на сетке $\{y_i\}_{i=\overline{0,n}}$ интерполяционный многочлен Лагранжа $L_n(y)$ для обратной функции $g(y)$

$$L_n(y) = \sum_{k=0}^{n} \frac{\omega(y)}{(y - y_k)\omega'(y_k)} \cdot x_k.$$

Тогда

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

что "легко" вычисляется (свободный член $L_n(y)$).

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 очевидно — метод секущих). Для случая $n = 2$ получить формулы.*

[*1)] Написать формулу наименьшего по модулю корня z_n

§4. Сходимость итерационных методов для уравнений $f(x) = 0$ с одним неизвестным

4.1 Достаточное условие существования и единственности решения. Сходимости метода простой итерации

Вернемся к итерационному уравнению (2)

$$x = \varphi(x)$$

в случае одной переменной. Для локализации корня применим к этому случаю *Теоремы T1 и T2:* Шар

$$\overline{K}(a;r) = \{x \in R^1 : |x - a| \le r\} = [a - r; a + r]$$

— отрезок $[a-r; a+r]$ числовой прямой. Для гладких отображений $\varphi(x)$ достаточное условие сжатия формулируется через ограничение роста модуля производной $|\varphi'(x)|$ в рассматриваемой области. Получим

Теорема 3. *Если $\varphi(x)$ — непрерывно-дифференцируема и $|\varphi'(x)| \le q < 1$ на отрезке $\overline{K}(a;r)$, а сам отрезок $\overline{K}(a;r)$ таков, что*

$$|\varphi(a) - a| \le (1 - q)r,$$

то:

1) Уравнение $x = \varphi(x)$ имеет на отрезке $\overline{K}(a;r)$ единственное решение $x^* \in \overline{K}(a;r)$.

2) Последовательность $x_{n+1} = \varphi(x_n)$, $x_0 \in \overline{K}(a;r)$ – дано (например: $x_0 = a$); сходится к неподвижной точке x^*, $x_n \underset{n\to\infty}{\to} x^*$.

Доказательство. Действительно, при сформулированных условиях $\varphi(x)$ — сжатие в шаре $\overline{K}(a;r)$ с коэффициентом $q < 1$ поскольку

$$|\varphi(x') - \varphi(x'')| = \left| \begin{matrix}\text{применяя}\\ \text{формулу}\\ \text{Лагранжа}\end{matrix} \right| = |\varphi'(\xi)||x' - x''| \le q|x' - x''|.$$

Таким образом выполнены все требования *Теоремы 2* $\Rightarrow$ существует и единственна неподвижная точка $x^* \in \overline{K}(a;r)$. Последовательность x_n при произвольном выборе $x_0 \in [a - r; a + r]$ сходится к ней $x_n \to x^*$ ■

Удобным практически способом локализации относительно x^* является проверка достаточных условий:

Теорема 4. *Если уравнение (2) имеет решение x^* и $\varphi(x)$ - непрерывно-дифференцируемая в окрестности корня x^* функция, причем $|\varphi'(x^*)| < 1$, то $\exists\ \varepsilon > 0$ такое, что на отрезке $\overline{K}(x^*, \varepsilon)$ уравнение (2) имеет единственное решение x^* и метод последовательных приближений (МПП) (3) сходится к x^* при произвольном $x_0 \in \overline{K}(x^*, \varepsilon)$.*

Действительно: Из непрерывности $\varphi'(x)$ на отрезке $\overline{K}(x^*, r) \Rightarrow \exists q \in (0, 1)$ и $\varepsilon > 0$ такие, что

$$|\varphi'(x)| \le q < 1 \quad \text{при} \quad x \in \overline{K}(x^*, \varepsilon).$$

Далее *Теорема 3* (для x^* выполнено условие $x^* - \varphi(x^*) = 0$) ■

4.2 Оценка погрешности метода последовательных приближений

Адаптируем к нашему случаю оценки (4) и (5) для погрешности МПП[*1)]. Напомним

$$\rho(x_n; x^*) = |x_n - x^*| = |\varphi(x_{n-1}) - \varphi(x^*)| \le q|x_{n-1} - x^*| \le \ldots \le q^n|x_0 - x^*|, \qquad (4')$$

МПП–метод первого порядка. Эта оценка позволяет утверждать не хуже, чем линейную сходимость МПП.
Оценка

$$|x_n - x^*| \le \frac{q^n}{1-q}|x_0 - x_n|, \qquad (5')$$

дает возможность эффективно оценить погрешность на n-ом шаге только через известные величины.

Для построения методов последовательных приближений, обладающих выше, чем линейной, сходимостью, нужны дополнительные требования на $\varphi(x)$.
Достаточно: *Если $\varphi(x)$ такова, что:*

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

то

$$x_{n+1} - x^* = \varphi(x_n) - \varphi(x^*) = \left| \begin{matrix} \textit{Формула Тейлора с} \\ \textit{центром в точке } x^* \end{matrix} \right| = \frac{(x_n - x^*)^p}{p!}\varphi^{(p)}(\xi); \quad \xi \in (x_n, x^*)$$

и при $|\varphi^{(p)}(x)| \le M_p$ получим

$$|x_{n+1} - x^*| \le \frac{M_p}{p!}|x_n - x^*|^p$$

т.е. $\{x_n\}$ интерационный процесс со сходимостью не ниже p-го порядка.
Теперь мы можем продолжить

$$|x_{n+1} - x^*| \le \frac{M_p}{p!}\left(\frac{M_p}{p!}|x_{n-1} - x^*|^p\right)^p \le \ldots \le \left(\begin{matrix} \text{продолжить и получить окончательную} \\ \text{формулу погрешности метода } p\text{-го} \\ \text{порядка} \end{matrix} \right). \qquad (6)$$

Замечания: 1) Формулы (4), (5) и (6) носят асимптотический характер и обеспечивают полученную сходимость итерационного метода в достаточно малой окрестности решения x^*.
2) Метод последовательного приближения, как и любой другой итерационный метод, выгодно отличается тем, что в нём не накапливается ошибка вычислений. Ошибка вычислений эквивалентна некоторому ухудшению очередного приближения, что отразится на числе итераций, но не на окончательной точности (если только итерационная последовательность остаётся на отрезке $\overline{K}(a, r)$).

[*1)] в R^1 своя метрика $\rho(x, y) = |x - y|$

4.3 Достаточные условия сходимости основных итерационных методов решения f(x)=0

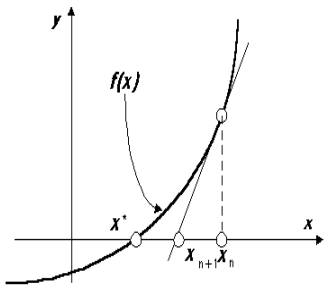
a) Метод релаксации. В методе релаксации

$$\varphi(x) = x + \tau f(x); \quad \varphi' = 1 + \tau f'; \quad |\varphi'(x^*)| \le 1 \Leftrightarrow$$
$$-1 < \varphi'(x^*) < 1; \quad \Leftrightarrow \quad -2 < \tau f'(x^*) < 0 \quad \left(\begin{matrix}\text{условия выбора}\\ \text{итерационного}\\ \text{параметра } \tau\end{matrix}\right). \tag{7}$$

Согласно *Теореме 4 метод релаксации* сходится при (7) и соответствующем выборе x_0.

b) Метод Ньютона. Ограничимся случаем простого корня, т.е. $f'(x^*) \neq 0$. Тогда

$$\varphi(x) = x - \frac{f(x)}{f'(x)}; \quad \varphi' = 1 - \frac{f'^2 - ff''}{f'^2} = \frac{ff''}{f'^2} \Rightarrow \varphi'(x^*) = 0 \quad (!)$$



Таким образом это

1) Метод второго порядка (не хуже). Квадратичная сходимость обеспечивается в некоторой окрестности корня x^*;

2) подавно имеет место *Теорема 4* при условии что $x_0 \in \overline{K}(a, r)$, в той области где $|\varphi'(x)| \le q < 1$.

3) На практике: после того, как уединен корень, выбирают x_0 так, чтобы $f(x_0)f'(x_0) > 0$. Выполнение такого условия дает одностороннюю сходимость метода последовательного приближения (см. рисунок).

4.4 Ускорение сходимости линейных итерационных методов

Используем метод Эйткена повышения порядка точности итерационных формул. Пусть итерационный метод имеет линейную сходимость и нам известно три последовательных расчета: x_{n-2}, x_{n-1}, x_n. Используем оценку (4) (ограничимся случаем односторонней сходимости)

$$x_n - x^* \approx q^n(x_0 - x^*) + O(q^{n+1}).$$

Для эффективной оценки q и $(x_0 - x^*)$ наши расчёты дают

$$\begin{aligned} x_n - x^* &= q^n(x_0 - x^*) + O(q^{n+1}) \\ x_{n-1} - x^* &= q^{n-1}(x_0 - x^*) + O(q^n) \\ x_{n-2} - x^* &= q^{n-2}(x_0 - x^*) + O(q^{n-1}) \end{aligned} \quad \Rightarrow$$

откуда

$$(x_{n-1} - x^*)^2 = (x_n - x^*)(x_{n-2} - x^*) \quad \Leftrightarrow$$
$$x_{n-1}^2 - 2x^* x_{n-1} + \cancel{x^{*2}} = x_n x_{n-2} - x^*(x_n + x_{n-1}) + \cancel{x^{*2}}$$

Тогда получим (в главном порядке по q)

$$x^* = \frac{x_{n-1}^2 - x_n x_{n-2}}{2x_{n-1} - x_n - x_{n-2}} = \frac{(x_{n-1} - x_n)^2 - x_n^2 + 2x_n x_{n-1} - x_n x_{n-2}}{2x_{n-1} - x_n - x_{n-2}} \Rightarrow$$

Окончательно мы получаем формулу $(n+1)$-го порядка точности

$$x^* = x_n + \frac{(x_{n-1} - x_n)^2}{2x_{n-1} - x_n - x_{n-2}} + O(q^{n+1}) \tag{8}$$

Полученная формула позволяет на очередном шаге итераций получить повышенную точность расчёта (если, конечно, не слишком велики накладные расходы при получении результата по формуле (8)).
Задание. *Получить оценку величины q.*

§5. Итерационные методы решения систем нелинейных уравнений

5.1 Постановка задачи. Каноническая форма одношагового итерационного метода

Напомним основное уравнение (1) для случая многих переменных

$$f_i(x_1, ..., x_m) = 0, i = 1, \ldots, m; \quad \Leftrightarrow \quad F = 0.$$

Каноническая форма записи *одношагового итерационного метода* такова:

$$\begin{cases} A_{k+1}\frac{\overset{(k+1)}{x} - \overset{(k)}{x}}{\tau_{k+1}} + F(\overset{(k)}{x}) = 0 \\ \overset{(0)}{x} = x_0. \end{cases} \tag{9}$$

Здесь τ_{k+1} - числовой итерационный параметр: A_{k+1} - невырожденная $\forall k$ матрица размерности $m \times m$; $detA_{k+1} \neq 0$: $k = 0, 1, 2, \ldots$

Очередное приближение $\overset{(k+1)}{x}$ ищется из решения системы линейных уравнений

$$A_{k+1}\overset{(k+1)}{x} = A_{k+1}\overset{(k)}{x} - \tau_{k+1} \cdot F(\overset{(k)}{x}) \quad \Leftrightarrow \quad \overset{(k+1)}{x} = \Phi(\overset{(k)}{x}).$$

Метод называется *явным*, если $A_{k+1} = E \quad \forall k$, т.е. в каждое i-е уравнение входит по одному неизвестному $\overset{(k+1)}{x_i}$. Метод называется *стационарным*, если $A_{k+1} = A, \quad \tau_{k+1} = \tau$ — не зависят от k.

5.2 Простейшие примеры одношаговых итерационных методов

а) *метод релаксации*

$$\frac{\overset{(k+1)}{x} - \overset{(k)}{x}}{\tau} + F(\overset{(k)}{x}) = 0 \quad \Leftrightarrow \quad \overset{(k+1)}{x} = \overset{(k)}{x} - \tau F(\overset{(k)}{x}) \tag{10}$$

явный стационарный метод.

б) *метод Ньютона (метод линеаризации).* Получается разложением $(1')$ в окрестности $\overset{(k)}{x}$ в ряд Тейлора с учетом лишь линейных относительно $dx \equiv \Delta x$ членов:

$$F(x) = 0 \quad \Leftrightarrow \quad F(\overset{(k)}{x}) + dF(\overset{(k)}{x}) + \frac{1}{2!}d^2F(\overset{(k)}{\xi}) = 0.$$

Удерживая лишь линейные относительно Δx члены получим

$$F_j(\overset{(k)}{x}) + \sum_1^n \frac{\partial F_j}{\partial x_i}(\overset{(k)}{x})(\overset{(k+1)}{x_i} - \overset{(k)}{x_i}) = 0$$

или

$$F'(\overset{(k)}{x})(\overset{(k+1)}{x} - \overset{(k)}{x}) + F(\overset{(k)}{x}) = 0.$$

Матрица Якоби $F'(\overset{(k)}{x})$ преобразования F считается невырожденной на каждом шаге итераций по k. (При рассмотрении модифицированного метода Ньютона итерационную матрицу фиксируют либо в начале алгоритма $F'(\overset{(0)}{x})$, либо на очередном шаге обновления $F'(\overset{(k_p)}{x})$.)

Итак, $detF'(x) = det(\frac{\partial F}{\partial x}) = det\|(\frac{\partial F_j}{\partial x_i})\| \neq 0$ в точке $\overset{(k)}{x} \quad \forall k \quad \Rightarrow$

$$\overset{(k+1)}{x} = \overset{(k)}{x} - \left(F'(\overset{(k)}{x})\right)^{-1} F(\overset{(k)}{x}) \equiv \Phi(\overset{(k)}{x}). \tag{11}$$

5.3 Сходимость метода Ньютона

Сформулируем достаточное условие сходимости метода Ньютона. Предположим, что $F(x)$ непрерывно-дифференцируемая функция и якобиан $detF'(x) \neq 0$ в некоторой окрестности точки x^*. Тогда

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

и

$$\frac{d\Phi}{dx} = E - \left\{\frac{d}{dx}(F'(x))^{-1}F(x) + (F'(x))^{-1}\frac{dF}{dx}\right\} = -\frac{d}{dx}((F'(x))^{-1})F(x).$$

Тем самым в точке x^* имеем $\frac{d\Phi}{dx}|_{x=x^*} = 0$, следовательно это метод второго порядка точности.

Пусть в пространстве x-ов введена какая-либо норма $\|x\|$. Поскольку $\Phi'(x)$- непрерывная функция своих аргументов x в некоторой окрестности т. x^*, то $\|\Phi'(x)\|$ — непрерывная числовая функция своих аргументов x_i в этой окрестности. Тогда существует достаточно малый шар $\overline{K}(x^*, \rho)$, где $\|\Phi'(x)\| \leq q < 1$. (Рассматривается норма матрицы Φ' согласованная с нормой x).

Рассмотрим этот шар $\overline{K}(x^*, \rho)$, и пусть $q = max_{x \in \overline{K}(x^*,\rho)}\|\Phi'(x)\| < 1$ (поскольку непрерывная функция на замкнутом ограниченном множестве достигает своего максимума).

В таком случае можно утверждать, что $\Phi(x)$ — сжатие в шаре $\overline{K}$ с коэффициентом q:

$$\Phi(x') - \Phi(x'') = \Phi'(\xi)(x' - x'') \Rightarrow$$
$$\|\Phi(x') - \Phi(x'')\| \leq ||\Phi'(\xi)|| \cdot \|(x' - x'')\| \leq q\|x' - x''\|$$

при этом $q < 1$.
Согласно *Теореме 2* корень x^* отделен и локализован в замкнутом шаре $\overline{K}(a, r) \subseteq \overline{K}(x^*, \rho)$ таком, что

$$\|a - \Phi(a)\| \leq (1 - q)r$$

В таком случае $\overset{(k)}{x} \to x^*$, если $\overset{(0)}{x} \in \overline{K}(a, r)$, например $\overset{(0)}{x} = a$.

Замечание: Оценим погрешность МПП прямо в терминах целевой функции F уравнения $(1')$. Имеем

$$F(\overset{(k)}{x}) = F(\overset{(k)}{x}) - F(x^*) = \frac{dF}{dx}(\overset{(k)}{\xi})(\overset{(k)}{x} - x^*) \Leftrightarrow$$
$$\|\overset{(k)}{x} - x^*\| \leq \|(\frac{dF}{dx}(\xi))^{-1}\| \cdot \|F(\overset{(k)}{x})\|.$$

При условии доступности оценки $\max_{x \in \bar{K}(a;r)} \|(\frac{dF}{dx})^{-1}\| = M$ получим

$$\|\overset{(k)}{x} - x^*\| \leq M\|F(\overset{(k)}{x})\|, \tag{12}$$

что связывает погрешность k-ой итерации и норму целевой функции на этой итерации.

ГЛАВА V

ЧИСЛЕННЫЕ МЕТОДЫ ЛИНЕЙНОЙ АЛГЕБРЫ

В нашем лекционном курсе мы остановимся на двух центральных проблемах численных методов линейной алгебры (ЛА). Это вопросы

I. Решение систем линейных алгебраических уравнений с невырожденной (квадратной) матрицей.

II. Нахождение собственных значений и собственных векторов для квадратных матриц — *алгебраическая проблема собственных значений.*

I. Решение систем линейных алгебраических уравнений

§1. Основные вычислительные задачи решения систем линейных алгебраических уравнений (СЛАУ)

1.1 Постановка задачи

В качестве основных вычислительных задач рассмотрим следующие три задачи:
1) Решение СЛАУ

$$Ax = f \qquad (\alpha)$$

с квадратной невырожденной матрицей $A_{n\times n}$, $detA \neq 0; x = \overrightarrow{x} = (x_1, ..., x_n)^T; f = \overrightarrow{f} \in R^n$; матрица A определяет отображение $A : R^n \Rightarrow R^n$.
2) Вычисление определителя матрицы $A = \|a^i_j\|^n_n = \|a_{ij}\|_{n\times n}$

$$\Delta = detA. \qquad (\beta)$$

3) Нахождение обратной матрицы A^{-1} для невырожденной квадратной матрицы A:

$$A^{-1}A = AA^{-1} = E. \qquad (\gamma)$$

Естественно, что приведенный перечень вопросов не охватывает все, и в том числе наиболее интересные практически, проблемы, связанные с решением СЛАУ.

Мы специально ограничиваемся рассмотрением невырожденной матрицы A, $detA \neq 0$, чтобы не привлекать другого подхода к понятию решения в случае, когда оно отсутствует, либо не является единственным [*1)].

1.2 Формальное решение. Устойчивость

Формальное решение задачи (α) строится по известным формулам Крамера

$$x = A^{-1}f; \quad x_i = \frac{\Delta_i}{\Delta}.$$

Формальное решение устойчиво, т.е. непрерывно зависит от входных данных A и $\vec{f}$. Действительно, варьируя $x = A^{-1}\vec{f}$, найдем [*2)]

$$\delta x = \delta(A^{-1}f) = \delta A^{-1}f + A^{-1}\delta f = \Big| \qquad \delta A^{-1} = -A^{-1}\delta A\, A^{-1}\Big| \quad \Rightarrow$$
$$\delta x = -A^{-1}\delta A\, A^{-1}f + A^{-1}\delta f = A^{-1}(\delta f - \delta A\, x). \qquad (*)$$

Таким образом $\|\delta x\| \to 0$ при $\|\delta f\|$ и $\|\delta A\| \to 0$.

1.3 Нормы

Напомним основные, используемые в R^n нормы
1) *Норма вектора* $\vec{x}$. Запишем разложение вектора $\vec{x}$ по базису $e = \{\vec{e}_i\}_n$:

$$\vec{x} = e\,X = (\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_n) \cdot (x_1, x_2, \ldots, x_n)^T,$$

Базисные векторы образуют строку e, а координаты вектора $\vec{x}$ — столбец X.

а) *евклидова норма* вектора

$$\|x\| = \sqrt{(\vec{x}, \vec{x})} = \sqrt{|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2};$$

б) l_p-норма (при $p = 2$ — норма Гильберта-Шмидта)

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

(для конечномерного случая $1/n$ можно перед суммой опустить).

в) c-норма (*равномерная* или *чебышевская* норма вектора x)

$$\|x\|_c = \sup_i |x_i| = \max_i |x_i| = \|x\|_\infty.$$

[*1)] при численных расчётах грань $detA \neq 0$ и $detA = 0$ достаточно условна

[*2)] учтём, что $\delta E \equiv 0 \equiv \delta(A^{-1}A) = \delta A^{-1}A + A^{-1}\delta A$

В R^n имеют место соотношения

$$\|x\|_1 \le \|x\|_2 \le \|x\|_c \le \sqrt{n}\,\|x\|_2 \le n\,\|x\|_1,$$

т.е. в R^n все эти нормы *эквивалентны* и сходимость в любой из них влечёт сходимость в остальных нормах.
Проверим, например:

$$\|x\|_c \le \sqrt{n}\,\|x\|_2.$$

Имеем

$$\sqrt{n}\,\|x\|_2 = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}|x_i|^2\right)^{1/2} = \left(\sum_{i=1}^{n}|x_i|^2\right)^{1/2} =$$

$$= \left(\max_i |x_i|^2 + \sum_{i\neq i_{max}}|x_i|^2\right)^{1/2} \le \max_i |x_i| = \|x_c\| \quad \blacksquare$$

Задача. *Доказать эквивалентность введенных норм.*
2) *Норма матрицы A.* Норма матрицы A, согласованная с нормой вектора $\overrightarrow{x}$ определяется следующим образом:

$$\|A\| = \sup_{\|x\|\neq 0}\frac{\|Ax\|}{\|x\|} = \left|\begin{array}{l}\text{в силу линейности}\\ \text{преобразования } A\\ \text{и свойств нормы}\end{array}\right| = \sup_{\|x\|\neq 0}\frac{\|A\frac{x}{\|x\|}\|x\|\|}{\|x\|} = \sup_{\|x\|=1}\|Ax\|.$$

Поскольку норма вектора — непрерывная функция его координат x_i, то на замкнутом, ограниченном множестве $\|x\| = 1$ она достигает своего наибольшего значения

$$\|A\| = \max_{\|x\|=1}\|Ax\| = \max_{\|x\|\neq 0}\frac{\|Ax\|}{\|x\|}.$$

Отсюда получаем, что

$$\|Ax\| \le \|A\|\cdot\|x\|.$$

Это условие согласования норм $\|x\|$ и $\|A\|$. Легко проверить, что введённая таким образом норма матрицы удовлетворяет неравенству

$$\|A+B\| \le \|A\| + \|B\|; \quad \|AB\| \le \|A\|\,\|B\|,$$

что и делает именно норму матрицы столь удобной в оценках.

Для квадратных матриц $A_{n\times n}$ наиболее употребительны следующие нормы:

$$\|A\|_c = \max_i\left(\sum_{j=1}^{n}|a_{ij}|\right) \qquad \|A\|_1 = \max_j\left(\sum_{i=1}^{n}|a_{ij}|\right)$$

$$\|A\|_M = n\cdot\max_{i,j}|a_{ij}| \qquad \|A\|_E = \max_i\left(\sum_i\sum_j|a_{ij}^2|\right)^{1/2}$$

$$\|A\|_S = \sqrt{\max_i\mu_i} = \max_i\nu_i$$

(где μ_i — собственные значения симметричной самосопряжённой матрицы (A^*A), $\nu_i = \sqrt{\mu_i}$). Первые две нормы не имеют специальных названий, $\|A\|_M$ называется

максимальной, $\|A\|_E$ — *сферической* или *евклидовой*, $\|A\|_S$ — *спектральной*. Эти нормы согласованы с нормами векторов в R^n норма $\|A\|_1$ согласована с нормой $\|x\|_1$, спектральная норма и сферическая — с $\|x\|_2$, максимальная норма $\|A\|_M$ — со всеми рассмотренными нормами векторов в R^n.

$$\|A\|_c \rightsquigarrow \|x\|_c; \quad \|A\|_1 \rightsquigarrow \|x\|_1; \quad \|A\|_E, \|A\|_S \rightsquigarrow \|x\|_2; \quad \|A\|_M \rightsquigarrow \|x\|_c, \|x\|_1, \|x\|_2.$$

Покажем согласованность $\|A\|_c$ и $\|x\|_c$:

$$\|A\,x\|_c = \max_i \left|\sum_j a_{ij}x_j\right| \le \max_i \sum_j |a_{ij}||x_j| \le \max_i \left(\underbrace{\max_j |x_j|}_{\|x\|_c} \sum_j |a_{ij}|\right) = \|x\|_c\,\|A\|_c \quad \blacksquare$$

Особенно часто используется евклидова норма $\|A\|_E$, поскольку она допускает сравнительно простую и наглядную интерпретацию и, что особенно важно, широкий класс геометрических преобразований сохраняет эту норму — это ортогональные преобразования:

$$U^TU = UU^T = E; \quad U^T = U^{-1}; \quad x \to Ux - \text{вращения и отражения относительно координатных плоскостей}$$

Задача. *Показать указанную согласованность норм.*

1.4 Обусловленность матрицы. Погрешности

Вернемся к анализу формулы (*) вариации решения x

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

1) Пусть матрица A известна точно ($\delta A = 0$) и погрешность решения связана лишь с погрешностью δf правой части, тогда

$$\delta x = A^{-1}\delta f \Rightarrow \|\delta x\| \le \|A^{-1}\| \cdot \|\delta f\|.$$

Из

$$f = Ax \Rightarrow \|f\| \le \|A\| \cdot \|x\|.$$

Перемножая полученные неравенства, найдем

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

или

$$\frac{\|\delta x\|}{\|x\|} \le \|A\| \cdot \|A^{-1}\| \frac{\|\delta f\|}{\|f\|} = CondA\,\frac{\|\delta f\|}{\|f\|}.$$

$CondA = \|A\|\|A^{-1}\|$ — число обусловленности матрицы A. $CondA \ge 1$ всегда (в любой норме[*1)]). Т.о. хорошо обусловлены матрицы с малым $CondA$, при этом относительная погрешность решения мала.

2) Пусть известно возмущение $\|\delta A\|$ матрицы A при условии, что правая часть f

[*1)] поскольку $E = A \cdot A^{-1} \Leftrightarrow \|E\| \le \|A\| \cdot \|A^{-1}\|$

задана точно. Тогда

$$\delta x = -A^{-1}\delta A\, x \Rightarrow \|\delta x\| \le \|A^{-1}\| \cdot \|\delta A\| \cdot \|x\|$$

или

$$\frac{\|\delta x\|}{\|x\|} \le \|A\| \cdot \|A^{-1}\| \frac{\|\delta A\|}{\|A\|} = CondA \frac{\|\delta A\|}{\|A\|}.$$

В общем случае, для малых возмущений матрицы A, когда $\|\delta A\| \ll \frac{1}{\|A^{-1}\|}$, и можно гарантировать существование обратной к $(A+\delta A)$ матрицы $(A+\delta A)^{-1}$, и получить оценку ее нормы через $\|\delta A\|, \|A\|, \|A^{-1}\|$ получим

$$\frac{\|\delta x\|}{\|x\|} \le \frac{CondA}{\underbrace{1 - CondA\frac{\|\delta A\|}{\|A\|}}_{\|A\|\cdot\|A^{-1}\| \ll 1}} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta f\|}{\|f\|} \right). \tag{2}$$

Замечания:

1) Анализ погрешностей по формуле (2) можно применять к случаю нахождения погрешности округления, как соответствующих возмущений A и f. Относительно громоздкими выкладками можно получить оценку относительной погрешности через машинное эпсилон ε_M для

$$\frac{\|\delta A\|}{\|A\|} = O(n\, 2^{-t}) = O(n\varepsilon_M)$$

здесь t–разрядность ЭВМ. Аналогичная оценка имеет место и для $\|\delta f\|/\|f\|$. Тогда, опираясь на (2), получим

$$\frac{\|\delta x\|}{\|x\|} = O(CondA\, n\, \varepsilon_M). \tag{3}$$

2) Априорное нахождение $CondA$ требует построения обратной к A матрицы и нахождения её нормы — это самостоятельная и весьма трудоёмкая задача.

Перейдем непосредственно к рассмотрению алгоритмов построения решения задачи (1) $Ax = f$.

§2. Метод Гаусса последовательного исключения неизвестных. LU - разложение

Численные методы решения СЛАУ (1) делятся на две большие группы: так называемые *прямые* и *итерационные* методы решения. *Прямые методы* дают решения СЛАУ за *конечное* число шагов. Они просты с алгебраической стороны и наиболее универсальны. Их основным недостатком является ограничение на порядок $n \sim 200$ матрицы системы уравнений (1), что связано с особенностью организации памяти доступных ЭВМ.

Итерационные методы используются, в основном, для решения СЛАУ специального (разреженного, слабозаполненного) вида с числом неизвестных $10^3 \div 10^5$ и более.

2.1 Формулы метода Гаусса

Одним из основных прямых методов решения СЛАУ является метод последовательного исключения неизвестных Гаусса. Он основан на возможности приведения исходной системы к эквивалентному представлению, когда относительно x решается задача с верхне-треугольной матрицей с единичной диагональю:

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

где $u_{ii} = 1, u_{ij} = 0$ при $j < i$.
Получение этой системы,т.е. построение матрицы U и вектора y составляют, так называемый, *прямой ход* метода исключения Гаусса. Дальнейшее решение системы $Ux = y$ — *обратный ход* метода исключения.

а) *Прямой ход исключения.* Опишем последовательно как он выполняется.
1-ый шаг. Пусть $a_{11} \neq 0$. Тогда, деля первое уравнение на a_{11}, получим

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

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

Комбинируя полученное уравнение с остальными уравнениями системы (1), исключая в них пееменную x_1, найдем:

$$\begin{array}{c} 0 \cdot x_1 + a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + ... + a_{2n}^{(1)}x_n = f_2^{(1)} \\ 0 \cdot x_1 + a_{32}^{(1)}x_2 + a_{33}^{(1)}x_3 + ... + a_{3n}^{(1)}x_n = f_3^{(1)} \\ \cdots \\ 0 \cdot x_1 + a_{n2}^{(1)}x_2 + a_{n3}^{(1)}x_3 + ... + a_{nn}^{(1)}x_n = f_n^{(1)} \end{array}$$

Для оставшихся уравнений (без x_1), повторим описанную процедуру:
S-ый шаг. После $(s-1)$ шагов исключения часть переменных исключена

$$\begin{array}{rcl} x_1 + u_{12}x_2 + ... + u_{12}x_n & = & y_1 \\ x_2 + u_{23}x_3 + ... + u_{2n}x_n & = & y_2 \\ \cdots & & \\ x_{s-1} + u_{s-1,s}x_s + ... + u_{s-1,n}x_n & = & y_{s-1} \end{array}$$

и мы получаем систему:

$$\left\{ \begin{array}{l} a_{s,s}^{(s-1)}x_s + a_{s,s+1}^{(s-1)}x_{s+1} + ... + a_{s,n}^{(s-1)} = f_s^{(s-1)} \\ ..... \\ a_{n,s}^{(s-1)}x_s + ... + a_{n,n}^{(s-1)}x_n = f_n^{(s)} \end{array} \right. \qquad (*)$$

Положим $a_{s,s}^{(s-1)} \neq 0$. Делим s-ое уравнение на $a_{s,s}^{(s-1)}$ и находим

$$x_s + u_{s,s+1}x_{s+1} + \cdots + u_{s,n}x_n = y_s \equiv \frac{f_s^{(s-1)}}{a_{s,s}^{(s-1)}}; \quad u_{s,j} = \frac{a_{s,j}^{(s-1)}}{a_{s,s}^{(s-1)}}.$$

Используем полученное уравнение. После умножения его на $a_{i,s}^{(s-1)}, \quad i = s+1, ..., n$ и вычитания из i-го уравнения исключаем x_s в системе $(*)$ из остальных уравнений.

Получим

$$\begin{aligned} x_s \quad + \qquad u_{s,s+1}x_{s+1} + \cdots + u_{s,n}x_n &= y_s \\ a^{(s)}_{s+1,s+1}x_{s+1} + \cdots + a^{(s)}_{s+1,u}x_n &= f^{(s)}_{s+1} \\ \cdots & \\ a^{(s)}_{n,s+1}x_{s+1} + \cdots + a^{(s)}_{n,n}x_n &= f^{(s)}_n, \end{aligned}$$

где

$$\begin{aligned} a^{(s)}_{ij} &= a^{(s-1)}_{i,j} - a^{(s-1)}_{i,s}u_{s,j} \quad ; \quad i,j = \overline{s+1,n} \\ f^{(s)}_i &= f^{(s-1)}_i - a^{(s-1)}_{i,s}y_s \quad ; \quad i,j = \overline{s+1,n} \end{aligned}$$

Таким образом прямой ход в методе Гаусса

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

осуществляется по формулам:

$$\begin{aligned} u_{s,j} &= \tfrac{a^{(s-1)}_{s,j}}{a^{(s-1)}_{s,s}}, \qquad s = 1,\ldots,n; \quad j = s+1,\ldots,n \\ a^{(s)}_{i,j} &= a^{(s-1)}_{i,j} - a^{(s-1)}_{i,s}u_{s,j}, \quad i,j = s+1,\ldots,n; \quad s = 1,\ldots,n-1 \end{aligned} \tag{4}$$

для матрицы и для правой части по формулам:

$$\begin{aligned} y_s &= \tfrac{f^{(s-1)}_s}{a^{(s-1)}_{s,s}}; \qquad s = 1,\ldots,n; \quad f^{(0)}_s = f_s \\ f^{(s)}_i &= f^{(s-1)}_i - a^{(s-1)}_{i,s}y_s; \quad i = s+1,\ldots,n; \quad s = 1,\ldots,n-1 \end{aligned} \tag{5}$$

б)*Обратный ход* метода Гаусса. Теперь решаем систему $Ux = y$ с верхнетреугольной матрицей, причём $u_{ii} = 1$:

$$\begin{cases} x_n = y_n \\ x_i = y_i - \sum\limits_{j=i+1}^{n} u_j x_j, \quad i = \overline{n-1,1} \end{cases} \tag{6}$$

Замечание. Формулы (4),(5) и (6) решают задачу (1). Число наиболее продолжительных арифметических действий — умножений-делений порядка $O(\frac{n^3}{3})$ и столько же сложений-вычитаний. Таким образом $O(\frac{2n^3}{3})$ арифметических действий необходимо для осуществления метода последовательного исключения неизвестных.

2.2 LU - разложение невырожденной матрицы

При реализации метода Гаусса на каждои шаге исключения мы полагали $a^{(s-1)}_{s,s} \neq 0$. Формулы (4) и (5) можно интерпритировать так, будто имеет место представление $f = Ly$ с нижней треугольной матрицей L

$$Ux = y = L^{-1}f \Leftrightarrow LUx = f \quad \text{т.е. } A = LU.$$

Это не случайно, однако само разложение мы получили по-другому (за одно и ответим на вопрос обоснования метода Гаусса).

Обозначим через Δ_i - главный угловой минор i-го порядка матрицы A. Тогда имеет место:

Теорема. *(LU-разложение) Пусть все угловые миноры матрицы A отличны от нуля, т.е. $\forall i\ \Delta_i \neq 0$, $i = 1, ...n$. Тогда матрицу A можно единственным способом представить в виде произведения $A = LU$, где L — невырожденная нижне-треугольная матрица; U — невырожленная верхне-треугольная матрица с единичной диагональю $u_{ii} = 1$.*

Доказательство: При $n = 1$ разложение очевидно

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

Пусть оно верно для $n = s - 1$, т. е.

$$A_{s-1} = L_{s-1} \cdot U_{s-1} \quad \text{и} \quad (U_{s-1})_{ii} = 1; \quad i = 1, \ldots, s-1.$$

Докажем, что наше утверждение верно для $n = s$. Для этого выделим в A_s удобную блочную структуру:

$$A_s = \left(\begin{array}{ccc|c} & & & a_{1,s} \\ & A_{s-1} & & \cdot \\ & & & \cdot \\ & & & a_{s-1,s} \\ \hline a_{s,1} & \cdots & a_{s,s-1} & a_{s,s} \end{array}\right) \quad \begin{array}{ll} \text{где} & \vec{b} = (a_{s,1}, \ldots, a_{s,s-1}); \\ \text{обозначено} & \vec{c} = (a_{1,s}, \ldots, a_{s-1,s})^T. \end{array}$$

Аналогичное разбиение на блоки выполним для матриц L_s и U_s. Вычислим $L_s U_s$ и потребуем

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

$$\left\{\begin{array}{lcl} L_{s-1}U_{s-1} & = & A_{s-1} \\ L_{s-1}\vec{u} & = & \vec{c} \\ \vec{l}U_{s-1} & = & \vec{b} \\ \vec{l}\vec{u} + l_{ss} & = & a_{ss} \end{array}\right. \Leftrightarrow \left\{\begin{array}{lcl} \vec{u} & = & L_{s-1}^{-1}\vec{c} \\ \vec{l} & = & \vec{b}(U_{s-1})^{-1} \\ l_{s,s} & = & a_{s,s} - \vec{l}\vec{u} \end{array}\right.$$

Мы использовали обозначения $\vec{l}$ и $\vec{u}$ для соответствующие векторов

$$\begin{array}{rcl} \vec{l} & = & (l_{s,1}, \ldots, l_{s,s-1}); \\ \vec{u} & = & (u_{1,s}, \ldots, u_{s-1,s})^T. \end{array}$$

Теперь покажем, что:

1)L_s - невырожденная матрица, (т.е. $l_{i,i} \neq 0$). Нам нужно показать, что $l_{s,s} \neq 0$. Остальные диагональные элементы матрицы L ненулевые по предположению индукции. Имеем

$$detA_s = \Delta_s = det(L_s U_s) = detL_{s-1} l_{s,s} \underbrace{detU_{s-1}}_{\equiv 1} \cdot 1 = detL_{s-1} \cdot l_{s,s} \neq 0.$$

Таким образом $l_{s,s} \neq 0$.

2) Разложение единственно (*от противного*). Пусть их два

$$A = \bar{L}\bar{U} = \tilde{L}\tilde{U} \Rightarrow \tilde{L}^{-1}\bar{L} = \tilde{U}\bar{U}^{-1}.$$

Здесь $\tilde{L}^{-1}\bar{L}$ — нижне-треугольная матрица, а $\tilde{U}\bar{U}^{-1}$ — верхне-треугольная матрица (с единичной диагональю). Таким образом $\tilde{L}^{-1}\bar{L}$ и $\tilde{U}\bar{U}^{-1}$ диагональные матрицы, но одна из них единиченая E

$$\tilde{L}^{-1}\bar{L} = E \Leftrightarrow \bar{L} = \tilde{L}; \quad \tilde{U}\bar{U}^{-1} = E \Leftrightarrow \tilde{U} = \bar{U} \quad \blacksquare$$

Замечания:
1) Как мы видим метод исключений Гаусса можно применять если все $\Delta_i \neq 0$. Известно, что если матрица A невырождена, $detA \neq 0$, то существует матрица перестановок P (не единственная!) такая, что PA (при перестановке строк A матрицы) имеет ненулевые главные миноры ($\Delta_i(PA) \neq 0$) и, следовательно, $PA = LU$ (единственным образом). В таком случае к системе

$$PA = Pf$$

и далее применим метод исключения Гаусса.
Один из способов реализации допустимой матрицы перестановок P — исключение с выбором главного элемента (т.е. элемента с максимальным модулем для исключения на соответствующем шаге) по строке или по столбцу (или по всей матрице). Такое исключение дает нужную перестановку уравнений. Окончательно

$$P = \prod_{(k,e)} P_{k,l}$$

где $P_{k,l}$ - матрица перестановки k, l строк (или столбцов).
2)**Задача.** Получить формулы *метода квадратного корня* (метода Холецкого). Пусть $A > 0; \quad A^T = A$, тогда

$$A = L\,L^T.$$

2.3 Вычисление определителя и обратной матрицы

Построенное LU-разложения для матрицы A позволяет решить вопрос о нахождении определителя и обратной матрицы для матрицы A.

Определитель матрицы. Обычно стандартные процедуры одновременно с построением решения СЛАУ вычисляют и определитель матрицы A. Пусть в процессе исключений найдено LU-разложение для A: $A = LU$, тогда

$$detA = det(LU) = detL \cdot detU = l_{11}....l_{nn}. \tag{7}$$

Если при исключении выполнялись перестановки, то

$$det(PA) = (-1)^{N(p)} detA$$

где $N(P)$ — число выполненных перестановок.
Если A вырожденная матрица, $detA = 0$, то на некотором, s-ом шаге исключения, $a_{s,i}^{(s-1)} = 0, i = \overline{s,n}$, что и завершает процесс исключения. При этом мы можем опредилить $rangA$ и построить базисные столбцы матрицы A.

Обращение матрицы. После получения LU-разложения для обращения матрицы A решают матричное уравнение

$$AX = E.$$

Решение X даёт A^{-1}. Относительно векторов $\vec{x}_k = (X)_k^i, \quad i = 1, \ldots, n$ — столбцов матрицы X имеем n систем

$$AX_k = E_k. \tag{8}$$

При этом для решения системы (8) разложение A строится один раз(!). Общее число мультипликативных действий $O(n^3)$ "всего в три раза больше (!)", чем для решения исходной системы линейных уравнений.

§3. Метод "прогонки" решения СЛАУ ленточного вида

3.1 LU-разложение ленточной матрицы

Частным, но важным с точки зрения приложений, является случай СЛАУ со специального вида матрицей:

$$a_{i,j} = 0, \qquad \text{если } |i - j| > k$$

т.е. матрица "ленточного" вида относительно главной диагонали, с широтой ленты $(2k+1)$ элемент. Любая матрица ленточная, но интересен случай, когда $(2k+1) \ll n$. Мы ограничимся рассмотрением случай $k = 1$, т.е. ширина ленты $2k + 1 = 3$. Это случай 3^{-x}-диагональной матрицы.
Удобства работ с ленточными матрицами объясняется прежде всего:

а) компактностью способа их хранения — требуется хранить не более $n * (2k + 1)$ элементов (даже меньше), а не n^2 как в обычном случае;

б) структурой LU-разложения. Имеет место

Теорема. *Если матрица A с шириной ленты $(2k+1)$ имеет LU-разложение*[*1)], *то L и U — соответствующие треугольные ленточные матрицы*

$$\begin{array}{l} l_{i,j} \neq 0, \quad j = i-k, \ldots, i \\ u_{i,j} \neq 0, \quad j = i, \ldots, i+k; \quad u_{i,i} = 1. \end{array}$$

Существенное замечание. При работе с ленточными матрицами крайне невыгодна перестановка уравнений, поскольку при этом увеличивается ширина ленты. Ограничимся рассмотрением случая 3-х-диагональной матрицы A, для которого реализация LU-разложения носит название *метода прогонки.*

3.2 Формулы прогонки

Рассмотрим СЛАУ $Ax = f$ с трехдиагональной матрицей A

$$A = \begin{pmatrix} b_1 & c_1 & 0 & 0 & \ldots & 0 \\ a_2 & b_2 & c_2 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & a_{n-1} & b_{n-1} & c_{n-1} \\ 0 & 0 & 0 & \ldots & a_n & b_n \end{pmatrix}$$

Главную и побочные диагонали матрицы обозначим b, a и c. Запишем СЛАУ $Ax = f$ в развернутом виде:

$$\begin{array}{rcll} a_i x_{i-1} + b_i x_i + c_i x_{i+1} & = & f_i, & i = 1, n \\ a_1 & = & 0 & \\ c_n & = & 0. & \end{array} \tag{9}$$

Построим формулы LU-разложения:

$$L = \begin{pmatrix} \beta_1 & 0 & 0 & 0 & \ldots & 0 \\ \alpha_2 & \beta_2 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \alpha_{n-1} & \beta_{n-1} & 0 \\ 0 & 0 & 0 & \ldots & \alpha_n & \beta_n \end{pmatrix}, \quad U = \begin{pmatrix} 1 & \gamma_1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & \gamma_2 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 & 1 & \gamma_{n-1} \\ 0 & 0 & 0 & \ldots & 0 & 1 \end{pmatrix}, \quad A = LU.$$

Поэлементно получаем:

$$\left\{ \begin{array}{rcl} \beta_1 \cdot 1 & = & b_1 \\ \beta_1 \cdot \gamma_1 & = & c_1 \end{array} \right. \quad \text{и} \quad \left\{ \begin{array}{rcl} \alpha_k \cdot 1 & = & a_k \\ \alpha_k \cdot \gamma_{k-1} + \beta_k \cdot 1 & = & b_k \\ \beta_k \cdot \gamma_k & = & c_k. \end{array} \right.$$

Формулы прогонки:

$$\begin{array}{rcll} \alpha_k & = & a_k; & k = 2, \ldots, n \\ \beta_1 & = & b_1 & \\ \gamma_1 & = & \frac{c_1}{\beta_1} & \\ \beta_k & = & b_k - \alpha_k \cdot \gamma_{k-1} = b_k - a_k \cdot \gamma_{k-1}; & k = 2, \ldots, n \\ \gamma_k & = & \frac{c_k}{\beta_k}: & k = 2, \ldots, (n-1) \end{array} \tag{10}$$

[*1)] в этом случае перестановки делать нельзя!

LU-разложение построено. Собственное решение (9) строим в два этапа:
а)*прямой ход* прогонки — находим y из $Ly = f$

$$\begin{array}{rcl} y_1 & = & \frac{f_1}{\beta_1} \\ y_k & = & \frac{(f_k - a_k \cdot y_{k-1})}{\beta_k} : \quad k = 2, \ldots, n. \end{array} \tag{11}$$

б)*обратный ход* — находим x из $Ux = y$

$$\begin{array}{rcl} x_n & = & y_n \\ x_k & = & y_k - \gamma_k \cdot x_{k+1} : \quad k = (n-1), \ldots, 1. \end{array} \tag{12}$$

Замечания:
1) Рассмотрим достаточные условия существования и единственности LU-разложения (10) — условие $\beta_k \neq 0$, $\forall k$. Покажем, что если выполнено условие диагонального преобладания элементов матрицы A, т.е. $|b_i| \geq |a_i| + |c_i|$ и $|a_i| \neq 0$, то $\beta_k \neq 0$, $\forall k$ (т.е. разложение (10) возможно и единственно). Еще раз подчеркнем, что мы не можем делать перестановок в матрице A. Заметим, что если $|\gamma_{k-1}| < 1$ при некотором k, то далее все $\beta_k \neq 0$ и все $|\gamma_k| < 1$ до γ_{n-1}. Действительно:

$$|\beta_k| = |b_k - a_k \cdot \gamma_{k-1}| \geq |b_k| - |a_k| \cdot |\gamma_{k-1}| \geq \underbrace{|a_k|(1 - |\gamma_{k-1}|)}_{>0} + \underbrace{|c_k|}_{\geq 0} > 0.$$

Теперь заметим

$$|\gamma_k| = \frac{|c_k|}{|b_k - a_k \gamma_{k-1}|} \leq \frac{|c_k|}{|b_k| - |a_k| \cdot |\gamma_{k-1}|} \leq \frac{|c_k|}{|a_k| + |c_k| - |a_k| \cdot |\gamma_{k-1}|} < 1$$

Поскольку по условию $|b_i| \geq |a_i| + |c_i| > 0$, а в дроби $\frac{|c_k|}{|a_k|+|c_k|-|a_k|\cdot|\gamma_{k-1}|}$ выражение $|a_k|(1 - |\gamma_{k-1}|) \neq 0$ ($|a_k| \neq 0$ по условию). Найдем $|\gamma_1|$ из (10), полагая $|\gamma_0| = 0$ (этот коэффициент не используется). Тогда $|\gamma_1| < 1$ ■
2) *Матричная прогонка.* В задаче разностной аппроксимации систем обыкновенных дифференциальных уравнений второго порядка мы сталкиваемся с ситуацией аналогичной LU-разложению (10). Приходится строить разложение трехдиагональной матрицы A блочной структуры, когда блоки $[a].[b]$ и $[c]$ сами являются $p \times p$-матрицами (p определяется числом уравнений в системе). Опуская формулы LU-разложения приведем аналог системы (10):

$$\begin{array}{rcl} [\beta_1][1] = [B_1] & \Leftrightarrow & [\beta_1] = [B_1] \\ [\beta_1][\gamma_1] = [C_1] & \Leftrightarrow & [\gamma_1] = [\beta_1]^{-1} C_1 \\ [\alpha_k][1] = [A_k] & \Leftrightarrow & [\alpha_k] = [A_k] \\ [\alpha_k][\gamma_{k-1}] + [\beta_k][1] = [B_k] & \Leftrightarrow & [\beta_k] = [B_k] - [A_k][\gamma_{k-1}] \\ [\beta_k][\gamma_k] = [C_k] & \Leftrightarrow & [\gamma_k] = [\beta_k]^{-1}[C_k]. \end{array} \tag{13}$$

Задача. *Записать для рассматриваемого случая аналог формул прогонки (10). Построить формулы решения прямого хода (14) и обратного — (15).*

§4. Итерационные методы решения СЛАУ

4.1 Одношаговые итерационные методы. Основные понятия

Одним из наиболее эффективных приемов решения СЛАУ (1)

$$Ax = f$$

высокого порядка, в частности, СЛАУ, возникающие при разностной аппроксимации дифференциальных уравнений (как правило с ленточными матрицами), являются *итерационные* методы.

Если для получения приближения решения (1) на очередной итерации (на очередном шаге итерационного процесса) используется лишь предыдущее значение x, то такой итерационный метод называется *одношаговым* (или двуслойным).

Мы ограничимся рассмотрением одношаговых итерационных методов, каноническая форма записи которых представляется в виде:

$$B_{n+1}\frac{x_{n+1}-x_n}{\tau_{n+1}} + Ax_n = f, \quad n = 0, 1, .. \tag{16}$$

$$x_0 = x^0.$$

Здесь B_{n+1} $(detB_{n+1} \neq 0, \forall n)$ и $\tau_{n+1} > 0$ — итерационные матрица и параметр. Основное внимание мы уделим *стационарным* итерационным методам, т.е. $B_{n+1} = B; \tau_{n+1} = \tau > 0$. Если $B \neq E$, то метод называется *неявным*. Точность итерационного метода характеризуется величиной нормы *погрешности* решения на n-ой итерации

$$z_n = x_n - x; \quad x_n = x + z_n; \tag{16}$$

где x — решение (1), z_n — погрешность n-ой итерации.

Поскольку (16) линейное относительно x уравнение, то погрешность z_n удовлетворяет однородному уравнению:

$$B_{n+1}\frac{(x+z_{n+1})-(x+x_n)}{\tau^{n+1}} + A(x+z_n) = f \Leftrightarrow B_{n+1}\frac{z_{n+1}-z_n}{\tau^{n+1}} + Az_n = 0. \tag{17}$$

Для неявного итерационного метода (16) естественно потребовать, чтобы решение задачи для x_{n+1}

$$B_{n+1}x_{n+1} = B_{n+1}x_n + \tau^{n+1}(f - Ax_n) \equiv F_n$$

требовало бы меньшего объема вычислений, чем прямое решение $Ax = f$.

Запишем (16) в форме *метода последовательных приближений* (МПП):

$$x_{n+1} = B_{n+1}^{-1}(B_{n+1} - \tau_{n+1}A)x_n + B_{n+1}^{-1}\tau_{n+1}f = \left|\begin{array}{l}\text{для стационарного}\\ \text{итерационного}\\ \text{метода}\end{array}\right| =$$

$$= B^{-1}(B - \tau A)x_n + \tau B^{-1}f \equiv Cx_n + g \tag{16*}$$

где $C = E - \tau B^{-1}A$ — матрица перехода к очередной итерации.

4.2 Представление основных (простейших) итерационных методов

Напомним еще раз, что мы рассматривали только *стационарные итерационные методы*
а)*метод релаксации*

$$\frac{x_{n+1}-x_n}{\tau}+Ax_n=f,\quad n=0,1,.. \tag{18}$$
$$x_0=x^0.$$

Итерационная матрица $B=E$ и $\tau>0$ — явный, стационарный метод. Если $\tau=\tau_{n+1}>0$, то метод называется *методом Ричардсона.* Запишем решение

$$\vec{x}_{n+1}=\vec{x}_n-\tau(A\vec{x}_n-\vec{f}). \tag{18*}$$

б) Группа итерационных методов:*метод Якоби, метод Зейделя, метод верхней релаксации.*
Все они основаны на специальном представлении матрицы A:

$$A=A_L+D+A_U,$$

где обозначено

$$(A_L)_{ij}=\begin{cases}a_{ij}, & j<i\\ 0, & j\geq i\end{cases}\quad (A_U)_{ij}=\begin{cases}0_{ij}, & j\leq i\\ a_{ij}, & j>i\end{cases}\quad D=diag(a_{11},...,a_{nn}).$$

1) *метод Якоби.* Роль итерационной матрицы B выполняет матрица D, $B=D$; $\tau=1$ (для канонической формы метода), тогда

$$D\vec{x}_{n+1}=D\vec{x}_n-(A\vec{x}_n-f). \tag{19}$$

Хотя метод и *неявный*, но, поскольку матрица D — диагональная, легко записать решение

$$(\vec{x}_{n+1})_i=(\vec{x}_n)_i-\frac{1}{a_{ii}}(A\vec{x}_n-\vec{f})_i,\quad i=\overline{1,N} \tag{19*}$$

$a_{ii}\neq 0$ по допущению поскольку мы считаем, что B — невырожденная матрица; N — размерность вектора $\vec{x}$.
2) *Метод Зейделя.* Полагаем $B=(A_L+D)$ — нижняя треугольная часть матрицы A. Канонический вариант метода получается при $\tau=1$, тогда:

$$(A_L+D)(\vec{x}_{n+1}-\vec{x}_n)=-(A\vec{x}_n-f), \tag{20}$$

Хотя метод Зейделя и неявный, но (A_L+D) легко обратима:

$$(\vec{x}_{n+1})_i=\frac{1}{a_{ii}}\left\{-(A_U\vec{x}_n-\vec{f})_i-\sum_{k=1}^{i-1}(A_L)_{ik}(\vec{x}_{n+1})_k\right\},\quad i=1,..,N. \tag{20*}$$

3)*Метод верхней релаксации.* Этот метод обобщает метод Зейделя

$$(D+\omega A_L)\frac{\vec{x}_{n+1}-\vec{x}_n}{\omega}+A\vec{x}_n=\vec{f},$$

Итерационная матрица $B = D + \omega A_L$ также нижняя треугольная матрица; параметр $\tau = \omega > 0$ в канонической форме метода.

$$(D + \omega A_L)\vec{x}_{n+1} = (D + \omega A_L)\vec{x}_n - \omega A\vec{x}_n + \omega\vec{f} = -\left\{ \underbrace{(\omega A - D - \omega A_L)}_{\text{верхн. треуг. матрица}} \vec{x}_n - \omega\vec{f} \right\}. \quad (21)$$

Решение дается формулой

$$(\vec{x}_{n+1})_i = \frac{1}{a_{ii}} \left\{ \left(-[\omega A_U + (\omega - 1)D]\vec{x}_n + \omega\vec{f} \right)_i - \sum_{k=1}^{i-1} (\omega A_L)_{ik} (\vec{x}_{n+1})_k \right\}, \quad i = 1, .., N. \quad (21^*)$$

4.3 Сходимость итерационых методов

1) Используем общую формулу (16*) метода последовательных приближений

$$x_n = Cx_{n-1} + g = \left| \begin{array}{c} C = E - \tau B^{-1}A \\ g = \tau B^{-1} f \end{array} \right| = C(Cx_{n-2} + g) + g = C^2 x_{n-1} + (E + C)g = \cdots =$$

$$= C^n x_0 + (E + C + ... + C^{n-1})g.$$

Это тождество и для сходимости $\{\vec{x}_n\}$ (16*) необходима и достаточна сходимость соответствующего степенного матричного ряда:

$$\sum_{n=0}^{\infty} C^n = E + C + C^2 + ... + C^n + ...$$

Сформулируем без доказательства теорему о сходимости матричных рядов:

$$\sum_{n=0}^{\infty} \alpha_n C^n = \alpha_0 E + \alpha_1 C + \alpha_2 C^2 + ... + \alpha_n C^n + ... \quad (*)$$

Для этого рассмотрим *производящий* ряд

$$\sum_{n=0}^{\infty} \alpha_n \lambda^n. \quad (**)$$

Теорема 1. *Для сходимости матричного степенного ряда (*) необходимо и достаточно чтобы все собственные значения матрицы C принадлежали области сходимости производящего ряда (**),*

$$\forall i; \lambda_i() \in (-R, R),$$

где $1/R = \overline{\lim_{n\to\infty}} |\alpha_n|^{1/n}$.

Теперь нетрудно сформулировать

Теорема 2. *Для сходимости метода последовательных приближений (16*) при*

произвольном выборе $\vec{x}_0$ *необходимо и достаточно, чтобы все собственные значения матрицы перехода* C *удовлетворяли бы условию* $|\lambda_i| < 1$, $\forall i$.
При этом, в силу необходимого условия сходимости ряда (*): $C^n \to 0$, а

$$\sum_{n=0}^{\infty} C^n = \left|{}^{\text{сходится}}_{\text{и даёт}}\right| = (E-C)^{-1} = (\tau B^{-1}A)^{-1} = A^{-1}\cdot(\tau B^{-1})^{-1} =$$

$$= A^{-1}B\frac{1}{\tau} \quad \Rightarrow \left(\sum_{n=0}^{\infty} C^n\right)\vec{g} \quad \Rightarrow A^{-1}\not{B}\frac{1}{\not{\tau}}(\not{\tau}\not{B}^{-1}f) = A^{-1}f = \vec{x}.$$

Замечания:
1) $S_n = \sum_{i=0}^{n} C^i$, если сходится, то к $(E-C)^{-1}$. Действительно

$$(E-C)\cdot S_n = (E + C + ... + C^n) - (C + ... + C^{n+1}) = E - C^{n+1} \Rightarrow E.$$

2) Имея возможность оценить собственные значения λ_i через норму матрицы C получаем достаточные условия сходимости метода последовательных приближений (16*). Напомним, для собственного значения λ справедливо

$$Cx = \lambda x \quad \Rightarrow \quad \|Cx\| = |\lambda|\cdot\|x\|,$$

но

$$\|Cx\| \le \|C\|\cdot\|x\| \quad \Leftrightarrow \quad |\lambda| \le \|C\|.$$

Достаточно выполнения в любой норме условия $\|C\| < 1$, что обеспечивает сходимость метода последовательных приближений (в соответствующей согласованной норме)

$$\begin{array}{lll} \text{в равномерной норме} & - & \max\limits_i(\sum\limits_j |c_{ij}|) < 1 \\ \text{в норме } l_1 & - & \max\limits_j(\sum\limits_i |c_{ij}|) < 1 \\ \text{в евклидовой норме} & - & \sum\limits_{ij} |c_{ij}|^2 < 1. \end{array} \qquad (22)$$

2)Дальнейшее изучение сходимости итерационных методов продолжим для случая, когда A — *симметричная, положительно определенная* матрица.
Напомним:

- Для вещественной матрицы A неравенство $A > 0$ означает:

$$(Ax, x) > 0, \quad \forall x \in \mathcal{H}, \quad x \neq 0.$$

Из неравенства $A > 0$ следует существование такого положительного $\delta > 0$, что $(Ax, x) \ge \delta\|x\|^2$, т.е. $A \ge \delta\cdot E$. Действительно:

1) если A симметричная матрица и $A > 0$, то все её собственные значения положительны, их можно упорядочить $0 < \lambda_1 \le \ldots \le \lambda_n$ и имеет место неравенство:

$$\lambda_1 E \le A \le \lambda_n E.$$

В качестве δ можно выбрать $\min\limits_i \lambda_i$.

2) если $A > 0$ несимметричная матрица, то

$$\forall x \in \mathcal{H}, \quad x \neq 0, \quad (Ax, x) = \frac{1}{2}((Ax, x) + (x, A^* \cdot x)) = \frac{1}{2} \cdot ((A + A^*)x, x) > 0.$$

Тем самым матрица $\frac{1}{2}(A + A^*) \equiv A_0$ — симметричная и положительно определенная матрица и в таком случае $\delta = \min\limits_i \lambda_i(A_0)$.

- Из неравенства
$$(Ax, x) \geq \delta \|x\|^2$$
следует существование обратной матрицы A^{-1} (отображение, задаваемое матрицей A — взаимнооднозначно):$A > 0 \quad \Rightarrow \exists\ A^{-1}$.

- Неравенство $A \geq 0$ означает $(Ax, x) \geq 0\ \forall x \in \mathcal{H}$ и A^{-1} может и не быть вовсе (у матрицы A невырожденное ядро $ker\ A$).

- Случай положительно определенной матрицы A позволяет ввести в $\mathcal{H}$ соответствующую A-энергетическую норму:
$$\|x\|_A = \sqrt{(Ax, x)}$$
и получить достаточное условие сходимости итерационного процесса.

Теорема 3. *(Самарского) Пусть $A > 0$ положительно определенная симметричная матрица. Параметр $\tau > 0$ и $B > 0$ таковы, что*

$$(B - \frac{\tau A}{2}) > 0,$$

тогда итерационный процесс (16) сходится в квадратичной метрике для любого $x_0 \in \mathcal{H}$.*

Доказательство.

1) Покажем сходимость итерационной последорвательности в энергетической A-норме. Для погрешности итерационного метода имеем итерационное уравнение (17*)

$$z_{n+1} = (E - \tau B^{-1} A) z_n; \quad Az_{n+1} = (A - \tau A B_{-1} A) z_n.$$

Откуда

$$\|z_{n+1}\|_A^2 = (Az_{n+1}, z_{n+1}) = (Az_n - \tau AB^{-1}Az_n, z_n - \tau B^{-1}Az_n) =$$
$$(Az_n, z_n) - \tau(z_n, AB^{-1}Az_n) - \tau(Az_n, B^{-1}Az_n) + \tau^2(B^{-1}Az_n, AB^{-1}Az_n);$$

т.к. $A^* = A$, то

$$(z_n, AB^{-1}Az_n) = (A^*z_n, B^{-1}Az_n) = (Az_n, B^{-1}Az_n)$$

т.е.

$$\|z_{n+1}\|_A^2 = \|z_n\|_A^2 - 2\tau(BB^{-1}Az_n, B^{-1}Az_n) + \tau^2(AB^{-1}Az_n, B^{-1}Az_n) =$$
$$\|z_n\|_A^2 - 2\tau((B - \frac{\tau A}{2})B^{-1}Az_n, B^{-1}Az_n) \Rightarrow \|z_{n+1}\|_A^2 \le \|z_n\|_A^2.$$

Поскольку

$$2\tau((B - \frac{\tau A}{2})B^{-1}Az_n, B^{-1}Az_n) \ge 0$$

напомним, что $B - \frac{\tau A}{2} > 0$.

Итак последовательность норм $\|z_n\|_A$ не возрастает и ограничена снизу (нулем) $\Rightarrow$ следовательно существует

$$\lim_{n\to\infty} \|z_n\|_A < \infty$$

(пока нам достаточно просто существование этого предела).

2) Далее, из условия $(B - \frac{\tau A}{2}) > 0 \Rightarrow \exists\delta > 0$ такое, что:

$$((B - \frac{\tau A}{2})B^{-1}Az_n, B^{-1}Az_n) \ge \delta\|B^{-1}Az_n\|^2$$

где $\|B^{-1}Az_n\|$ — в среднеквадратичная норма. Тогда

$$\|z_{n+1}\|_A^2 - \|z_n\|_A^2 + 2\tau((B - \frac{\tau A}{2})B^{-1}Az_n, B^{-1}Az_n) = 0.$$

В этом тождестве заменим $2\tau((B - \frac{\tau A}{2})B^{-1}Az_n, B^{-1}Az_n)$ на меньшее. Получим

$$\|z_{n+1}\|_A^2 - \|z_n\|_A^2 + \delta\|B^{-1}Az_n\|^2 \le 0.$$

Поскольку последовательность A-норм z_n сходится, то $\|z_{n+1}\|_A^2 - \|z_n\|_A^2 \to 0$ при $n \to 0$. Таким образом

$$\lim_{n\to\infty} \|B^{-1}Az_n\| = 0.$$

Нами установлена среднеквадратичная сходимость последовательности $w_n = (B^{-1}Az_n)$.

Но поскольку A — положительно определённая матрица, то $\exists A^{-1}$ и $z_n = A^{-1}Bw_n$, причем

$$\|z_n\| \le \|A^{-1}B\| \cdot \|w_n\| \Rightarrow \|z_n\| \to 0, \quad n \to \infty$$

что и требовалось доказать ■

Замечания:

1) Сравнительно несложно показать, что имеет место сходимость итерационной последовательности $\{z_n\}$ в A-энергетической норме

$$\|z_n\|_A \to 0.$$

Сходимость именно к 0, т.е. эти нормы эквивалентны.

2) В A-норме сходимость первого порядка

$$\|z_{n+1}\|_A \le q\|z_n\|_A,$$

где

$$q = \sqrt{1 - \frac{2\tau\bar{\lambda}\tilde{\lambda}}{\|B\|^2}}, \quad \bar{\lambda} = \min_k \lambda_k(A), \quad \tilde{\lambda} = \min_k \lambda_k\left(\frac{B + B^*}{2} - \frac{\tau A}{2}\right). \tag{23}$$

4.4 Достаточные условия сходимости простейших итерационных методов

Применим достаточные условия *Теоремы3* к анализу простейших итерационных методов.
а) метод релаксации:

$$B = E; \quad B - \frac{\tau A}{2} = E - \frac{\tau A}{2} > 0$$

Этому неравенству можно удовлетворить выбором параметра τ. Напомним, из неравенства

$$\|A\| = \sup_{\|x\| \neq 0} \frac{(Ax, x)}{(x, x)} \Rightarrow (Ax, x) \leq \|A\| \cdot (x, x)$$

тем самым $A \leq \|A\| \cdot E$ или $\frac{A}{\|A\|} \leq E$. Получим

$$E - \frac{\tau A}{2} \geq \frac{A}{\|A\|} - \frac{\tau A}{2} = (\frac{1}{\|A\|} - \frac{\tau}{2})A > 0$$

таким образом:

$$\frac{1}{\|A\|} - \frac{\tau}{2} > 0 \quad \Rightarrow \quad 0 < \tau < \frac{2}{\|A\|}. \tag{24}$$

б) метод верхней релаксации:

$$B = D + \omega A_L; \quad \tau = \omega.$$

$$B - \frac{\tau A}{2} = D + \omega A_L - \frac{\omega}{2}(A_L + D + A\omega) = (1 - \frac{\omega}{2})D + \frac{\omega}{2}(A_L - A_U).$$

Положительная определённость матрицы $B - \frac{\tau A}{2}$ означает:

$$(B - \frac{\tau A}{2}x, x) = (1 - \frac{\omega}{2})(Dx, x) + \frac{\omega}{2}((A_L x, x) - (A_U x, x)) = (1 - \frac{\omega}{2})(Dx, x) > 0.$$

D силу симмтрии матрицы A: $((A_L x, x) - (A_U x, x)) \equiv 0$, поскольку $(A_L x, x) = (x, A_U^* x)$. Итак

$$(1 - \frac{\omega}{2})D > 0.$$

Из условия $A > 0$ следует, что матрица $D = diag\,(a_{11}, \ldots, a_{nn}) > 0$. Действительно, возьмём в качестве вектора $\vec{x}$ базисный вектор $\vec{e}_i$:

$$\vec{x} = \vec{e}_i = \underbrace{(0, ..., 1, ..., 0)^T}_{1 \text{ на } i\text{-ом месте}}; \quad (Ax, x) = a_{ii} = (Dx, x) > 0, \quad \text{т.е. все } a_{ii} > 0.$$

Таким образом для произвольного вектора $\vec{x}$

$$(Dx, x) = a_{ii} x_i^2 > 0.$$

Окончательно

$$\left(1 - \frac{\omega}{2}\right) > 0 \quad \Rightarrow \quad 0 < \omega < 2. \tag{25,26}$$

(В частности при $\omega = 1$ обеспечена сходимость метода Зейделя).
в) метод Якоби: $B = D$ и τ =1

$$B - \frac{\tau A}{2} = D - \frac{A}{2} > 0, \quad \Rightarrow \quad A < 2D. \tag{27}$$

Сформулируем достаточные условия сходимости метода Якоби
Теорема 4. *Если A симметричная положительно определённая матрица с диагональным преобладанием*

$$a_{ii} > \sum_{j \neq i} |a_{ij}|, \tag{28}$$

то метод Якоби сходится (в среднеквадраитичной метрике).
Действительно, покажем, что в таком случае выполнено неравенство (27):

$$(Ax, x) = \sum_{ij} a_{ij} x_i x_j \leq \sum_{ij} |a_{ij}| |x_i| |x_j| \leq \left| \begin{matrix} \text{учтём, что} \\ |x_i||x_j| \leq \frac{|x_i|^2 + |x_j|^2}{2} \end{matrix} \right| \leq$$

$$\leq \frac{1}{2} \left(\sum_{ij} |a_{ij}| |x_i|^2 + \sum_{ij} |a_{ij}| |x_j|^2 \right) = \left| \begin{matrix} \text{в силу} \\ \text{симметрии} \\ a_{ij} = a_{ji} \end{matrix} \right| =$$

$$= \sum_i |x_i|^2 \left(a_{ii} + \sum_{j \neq i} |a_{ij}| \right) < \sum_i |x_i|^2 2 a_{ii} = 2(Dx, x).$$

Таким образом $A < 2D$ что и требовавлось доказать ■

II. Алгебраическая проблема собственных значений

§1. Собственные значения (с.з.) и собственные векторы (с.в.) квадратной матрицы. Прямые методы

1.1 Основные понятия

Напомним:*Ненулевой вектор $\vec{x} \neq 0$ называется собственным вектором матрицы A, отвечающим собственному значению λ, если*

$$Ax = \lambda x, \quad x \neq 0. \tag{1}$$

В дальнейших рассуждениях мы будем считать, что $\overrightarrow{x} \in C_n$, A–квадратная комплексная матрица, задающая отображение$A : C_n \Rightarrow C_n$; $\lambda \in C$. Хотя, как правило, матрица A, $\overrightarrow{x}$ и λ — будут вещественны (в наших приложениях).
Необходимое и достаточное условие нетривиальной разрешимости (1):

$$\det(A - \lambda E) = 0. \tag{2}$$

Многочлен

$$p(\lambda) = P_n(\lambda) = \det(A - \lambda E)$$

называется характеристическим многочленом матрицы A.
Будем сразу же предполагать, что:

- Собственные векторы $\vec{x}$ нормированы, т. е. $\|\overrightarrow{x}\| = \sqrt{(x,x)} = 1$;
- Известно, что если *все* собственные значения матрицы A простые (не кратные), то она имеет n линейно независимых (ЛНЗ) собственных векторов, т. е. в этом случае существует базис в C_n из собственных векторов матрицы A.
- Если среди собственных значений есть кратные, то

  1) Собственные вектора, отвечающие различным (не комплексно сопряженным) собственным значения — линейно независимы;

  2) Собственных векторов для λ_i кратности α_i может и не быть α_i штук;

- Поскольку

  $$\overline{\det(A - \lambda E)} = \det\left(\overline{A - \lambda E}\right) = \det\left(\overline{A - \lambda E}\right)^T = \det\left(\overline{A}^T - \overline{\lambda} E\right) = \det\left(\overline{A^T} - \overline{\lambda} E\right),$$

  то, если λ — собственное значение матрицы A, то $\overline{\lambda}$ — собственное значение сопряжённой матрицы $\left(\overline{A^T}\right) \equiv A^*$.

- Собственные векторы сопряженных матриц, отвечающие различным (не комплексно сопряженным) собственным значениям — *ортогональны.*
- У эрмитовских матриц ($A^* = A$) все собственные значения *вещественны*, а собственные векторы образуют после ортогонализации *ортонормированную систему* (ОНС) и если матрица $A > 0$ (положительно определенная), то существует базис из собственных векторов матрицы A.

1.2 Устойчивость невырожденной задачи нахождения собственных векторов и собственных значений

Пусть собственные векторы матрицы A образуют базис в C_n и данное собственное значение λ_k — простое. Тогда возмущенная погрешностями задача (1) имеет вид:

$$(A + \delta A)(x_k + \delta x_k) = (\lambda_k + \delta\lambda_k)(x_k + \delta x_k).$$

Линеаризуя по возмущениям δA, $\delta\lambda_k$, $\delta\overrightarrow{x_k}$ и учтя, что $Ax_k = \lambda_k x_k$, найдем:

$$A\ \delta x_k + \delta A\ x_k = \lambda_k\ \delta x_k + \delta\lambda_k\ x_k. \tag{$*$}$$

Поскольку собственные векторы нормирваны $\|x_k\|^2 = (x_k, x_k) = 1$, то варьируя это равенство найдем $(\delta x_k, x_k) = 0$ (δx_k и x_k ортогональны.)
В таком случае в разложении δx_k по невозмущенному базису $\{x_i\}$ коэффициент $\alpha_{kk} = 0$, имеем

$$\delta x_k = \sum_{i=1}^{n}{}' \alpha_{ki} x_i,$$

Штрих у суммы означает, что $i \notin defind\{k\} = \{k\}$. Тогда

$$A \sum_{i=1}^{n}{}' \alpha_{ki} \overrightarrow{x_i} + \delta A \ \vec{x}_k = \lambda_k \sum_{i=1}^{n}{}' \alpha_{ki} \overrightarrow{x_i} + \delta\lambda_k \ \vec{x}_k. \qquad (**)$$

Теперь умножим (**) скалярно (т. е. справа) на собственный вектор y_l сопряжённой матрицы A^*, получим:
1) $l = k$, y_k — собственный вектор для $\overline{\lambda_k}$:

$$\overbrace{\left(A \sum_{i\neq k}^{n}{}' \alpha_{ki}\overrightarrow{x_i}, y_k\right)}^{=0 \text{ ибо } \perp y_k} + \overbrace{\left(A \sum_{i=1}^{n}{}' \alpha_{ki}\overrightarrow{x_i}, y_k\right)}^{\equiv 0} + (\delta A \ x_k, y_k) = \left(\lambda_k \sum_{i=1}^{n}{}' \alpha_{ki}\overrightarrow{x_i}, y_k\right) + \delta\lambda_k (x_k, y_k).$$

$$\Updownarrow$$

$$(\delta A \ x_k, y_k) = \delta\lambda_k (x_k, y_k).$$

Или

$$|\delta\lambda_k| \le \frac{|(\delta A \ x_k, y_k)|}{|(\ x_k, y_k)|} \le \frac{\|\delta A \ x_k\| \ \ \|y_k\|}{|(\ x_k, y_k)|} \le \frac{\|\delta A \ \| \ \ \|x_k\| \ \ \|y_k\|}{|(\ x_k, y_k)|} \le$$

$$\max_{i,j} \ |\delta a_{i,j}| \ \underbrace{\frac{\sqrt{(x_k, x_k)\,(y_k, y_k)}}{|(x_k, y_k)|}}_{\varkappa_{k,k}} \equiv \varkappa_{k,k} \max_{i,j} \ |\delta a_{i,j}|. \qquad (3)$$

здесь $\varkappa_{k,k}$ — k-ый главный коэффициент перекоса матрицы A.
2) Аналогично $l \neq k$:

$$\overbrace{\left(\sum_{i=1}^{n}{}' \alpha_{ki}\lambda_i \overrightarrow{x_i}, y_l\right)}^{\text{остается лишь } i=l} + (\delta A \ x_k, y_l) = \left(\lambda_k \sum_{i=1}^{n}{}' \alpha_{ki}\overrightarrow{x_i}, y_l\right) + \overbrace{\delta\lambda_k (\ x_k, y_l)}^{\equiv 0 \ x_k \perp y_k}$$

$$\alpha_{kl}\lambda_l \ \ (x_l, y_l) + (\delta A \ x_k, y_l) = \lambda_k \alpha_{kl} \ \ (x_l, y_l).$$

Откуда получаем

$$\alpha_{lk} (\lambda_k - \lambda_l) (x_l, y_l) = (\delta A \ x_k, y_l).$$

Теперь мы можем получить оценку коэффициентов α_{kl}

$$|\alpha_{kl}| \le \frac{|(\delta A \ x_k, y_l)|}{|(\lambda_k - \lambda_l)| \ \ |(x_l, y_l)|} \le \frac{\max_{i,j} |\delta a_{i,j}|}{|(\lambda_k - \lambda_l)|} \frac{\sqrt{(x_k, x_k)\,(y_l, y_l)}}{|(x_l, y_l)|} =$$

$$= \frac{\varkappa_{k,l}}{|(\lambda_k - \lambda_l)|} \max_{i,j} \ |\delta a_{i,j}|. \qquad (4)$$

Итак:

1) собственное значение λ_k матрицы A устойчиво относительно возмущений матрицы, если соответствующий ему коэффициент перекоса $\varkappa_{k,k}$ мал;

2) Для устойчивости собственных векторов относительно возмущений матрицы A необходимо, чтобы *все* $\varkappa_{k,l}$ были малы.

3) Для эрмитовских матриц $A^* = A$, $x_i = y_i$ и все $\varkappa_{k,l} = 1$. Тем самым задача (1) нахождения собственных значений и собственных векторов устойчива относительно возмущений входных данных δA.

1.3 Вычисление собственных значений (метод интерполяции)

Для нахождения собственных значений матрицы A используют её *характеристический многочлен*

$$p_n(\lambda) \equiv \det(A - \lambda E).$$

Далее можно любыми методами искать корни этого многочлена, т. е. решение уравнение (2)

$$p_n(\lambda) = 0.$$

Из общих соображений удобен *метод парабол*, поскольку он может обеспечить сходимость к комплексному корню характеристического уравнения (2) при действительном начальном приближении.

Как строить $p(\lambda)$? Естественно представить $p(\lambda)$ в виде интерполяционного многочлена, используя сетку значений $\{\lambda_i\}_{i=\overline{0,n}}$ (здесь λ_i — узел интерполяционной сетки, а не собственное значение матрицы A).

Сетку $\{\lambda_i\}$ выгодно брать на интервале $[-\|A\|, \|A\|]$ поскольку $\forall i \ |\lambda| \le \|A\|$. Тогда

$$p_n(\lambda) = \det(A - \lambda E) = L_n(\lambda) = \left| \begin{array}{c} \text{точное} \\ \text{равенство ибо} \\ \text{это полином} \\ \text{порядка } n \end{array} \right| =$$

$$= \sum_{k=0}^{n} \overbrace{p_n(\lambda_k)}^{=\det(A-\lambda_k E)} \frac{\omega(\lambda)}{\omega'(\lambda_k)(\lambda - \lambda_k)}, \qquad (5)$$

где $\omega(\lambda) = \prod_{i=0}^{n} (\lambda - \lambda_i)$.

Вычисления по формуле (5) требуют $O(\frac{2}{3}n^4)$ действий.

В практике вычислений предпочтительнее следующая тактика поведения:
Известно, что матрицу общего вида преобразованием *подобия* P можно привести к 3-х-диагональной матрице, т .е. $\exists P, \quad \det P \neq 0$ такая, что

$$B = P^{-1}AP$$

3-х-диагональная матрица. Тогда, если собственные значения и собственные векторы матрицы B известны, то

$$By = \alpha y \quad \Longleftrightarrow \quad P^{-1}APy = \alpha y \quad \Longleftrightarrow \quad A(Py) = \alpha(Py)$$

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 строиться через собственный вектор y_i

$$x_i = Py_i. \tag{6}$$

Для нахождения собственных значений матрицы B, т.е. корней характеристического многочлена $P_B(\alpha) = 0$, для *метода интерполяции* необходимо вычисление определителя $\det(B - \alpha E)$ в узлах сетки $\{\alpha_i\}$. Эти вычисления можно реализовать по экономичной расчётной схеме, что весьма важно. Именно

$$D_m(\alpha) = \begin{vmatrix} & & & & & & 0 \\ & D_{m-2} & & & & & \cdot \\ & & & & & & \cdot \\ & & & & & & 0 \\ \hline & & b_{m-1,m-2} & & b_{m-1,m-1} - \alpha & & b_{m-1,m} \\ \hline & & & & b_{m,m-1} & & b_{m,m} - \alpha \end{vmatrix} =$$

$$= (b_{m,m} - \alpha) \cdot D_{m-1}(\alpha) - b_{m,m-1} \cdot b_{m-1,m} \cdot D_{m-2}(\alpha).$$

Итак, мы получили рекурентные формулы вычисления характеристического многочлена 3-x –диагональной матрицы

$$\begin{cases} D_m(\alpha) = (b_{m,m} - \alpha) D_{m-1} - b_{m,m-1} \cdot b_{m-1,m} \cdot D_{m-2}(\alpha) \\ D_{-1} = 0; \quad D_0 = 1 \end{cases} \tag{7}$$

Задача. *Убедиться, что (7) дает правильную рекуррентную формулу для* $D_m(\alpha)$.
Однократное вычисление $P_B(\alpha_i) \equiv D_n(\alpha_i)$ требует $O(5n)$ действий, следовательно для десяти итераций метода парабол $\sim O(50n^2)$.

1.4 Нахождение собственных векторов (метод обратной итерации)

Если собственное значение λ_k матрицы A известно, то соответствующие ему собственные векторы $\vec{x}_k$ находятся из решения совместной СЛАУ (1). Однако:

-на решении СЛАУ с вырожденной матрицей мы не останавлилвались;

-особенно хорошего результата от непосредственного применения методов предыдущей части к решению (1) ожидать нельзя, ибо поскольку (в силу различного рода погрешностей) λ_k известно неточно $\tilde{\lambda}_k$, то формально $\det\left(A - \tilde{\lambda}_k E\right) \neq 0$. Тем самым её решение должно тождественно равняться нулю $\tilde{x}_k \equiv 0$.

Поэтому для решения задачи (1) используют другие соображения:

Метод обратной итерации: Выберем прозвольно вектор x_0 и рассмотрим СЛАУ

$$\left(A - \tilde{\lambda}_k E\right) x = x_0. \qquad (*)$$

Поскольку $det\left(A - \tilde{\lambda}_k E\right) \neq 0$, то у СЛАУ (*) существует единственное решение x . Покажем, что это решение "почти" собственный вектор x_k.

Пусть, как и прежде, A имеет n линейно независимых собственных векторов $\{x_i\}_{i=\overline{1,n}}$ — базис. Пусть λ_k — простое собственное значение (нам дано!). Тогда решение x и вектор x_0 можно разложить по базису из собственных векторов $\{x_i\}$

$$x = \sum_{i=1}^{n} \alpha_i x_i, \qquad x_0 = \sum_{i=1}^{n} \beta_i x_i.$$

Найдем α_i

$$Ax - \tilde{\lambda}_k x = \sum_i \alpha_i \lambda_i x_i - \sum_i \tilde{\lambda}_k \alpha_i x_i = \sum_i \beta_i x_i.$$

В силу единственности разложения по выбранному базису

$$\alpha_i \left(\lambda_i - \tilde{\lambda}_k\right) = \beta_i \quad \Leftrightarrow \quad \alpha_i = \frac{\beta_i}{\lambda_i - \tilde{\lambda}_k} \qquad (**)$$

Итак мы получили, что все коэффициенты α_i в разложении вектора x относительно малы, кроме α_k (в знаменателе дроби (**) стоит при $i = k$ малая величина).

Найденный вектор $\vec{x}$ нужно обязательно нормировать и провести несколько итераций

$$\begin{cases} \left(A - \tilde{\lambda}_k E\right) \overset{(s)}{\vec{x}}_k = \overset{(s-1)}{\vec{x}}_k ; \quad \overset{(0)}{\vec{x}}_k = \vec{x}_0 \\ \qquad\qquad |\overset{(k)}{\vec{x}}_k| = 1. \end{cases} \qquad (8)$$

Замечания:

-Обычно достаточно $2^x \div 3^x$ итераций (с нормировкой !) для достижения приемлемой точности;

-В случае кратных собственных значений, если собственные векторы образуют базис, то $\vec{x} \in Lin\,(\overset{(k)}{\vec{x}}_1, \ldots, \overset{(k)}{\vec{x}}_{\alpha_k})$ и для линейно независимых начальных значений $\{\vec{x}_0\}_{\overline{1,\alpha_k}}$ найдём α_k линейно независимых собственных векторов в этой оболочке.

§2. Итерационный метод вращений (Якоби) нахождения собственных векторов и собственных значений симметричной матрицы

2.1 Редукция

Известно, что для действительной симметричной матрицы $A^T = A$ существует ортогональное преобразование U такое, что

$$U^T A U = diag\,(\lambda_1, ..., \lambda_n) = \Lambda.$$

(частный случай преобразования подобия $U^{-1} = U^T$).

У матрицы Λ собственными векторами являются базисные векторы $\vec{e}_i = \{0, 0, ..., 1, 0, ..., 0\}$ ибо

$$\Lambda\vec{e}_i = \lambda_i\vec{e}_i \quad \Leftrightarrow \quad \left(U^T A U\right)\vec{e}_i = \lambda_i\vec{e}_i \quad \Leftrightarrow \quad A\left(U\vec{e}_i\right) = \lambda_i\left(U\vec{e}_i\right).$$

Таким образом λ_i — собственные значения матрицы A, а

$$\vec{x}_i = U\vec{e}_i$$

собственные векторы исходной матрицы A, отвечающие собственному значению λ_i.

2.2 Построение матрицы вращения U

В методе Якоби матрица U строится итерационно, через последовательность элементарных вращений. Рассмотрим вращение в плоскости (Okl)

$$U_{k,l}(\varphi) = \begin{pmatrix} 1 & & 0 & & & & & & \\ & \ddots & & & & 0 & & & 0 \\ 0 & & 1 & & & & & & \\ & & & \cos\varphi & & & -\sin\varphi & & \\ & & & & 1 & & 0 & & \\ & 0 & & & & \ddots & & & 0 \\ & & & & 0 & & 1 & & \\ & & & \sin\varphi & & & \cos\varphi & & \\ & & & & & & & 1 & \\ & 0 & & & & 0 & & & \ddots \\ & & & & & & & & & 1 \end{pmatrix}, \quad \tilde{A} = U_{kl}^T A U_{kl}.$$

И определим φ из условия $\left(\tilde{A}\right)_{kl} = \left(\tilde{A}\right)_{lk} = 0$. Найдем $\left(\tilde{A}\right)_{kl}$:

1) Матрица $B = AU_{kl}$ отличается от A двумя столбцами k-ым и l-ым (все остальные столбцы в U единичные):

$$b_{ik} = \sum_j a_{ij}\left(U_{kl}\right)_{jk} = a_{ik}\cos\varphi + a_{il}\sin\varphi$$

$$b_{il} = \sum_j a_{ij}(U_{kl})_{jl} = a_{ik}\,(-\sin\varphi) + a_{il}\cos\varphi.$$

2) Матрица $\tilde{A} = U^T B$ — отличается от B двумя строками: k-ой и l-ой:

$$\left(\tilde{A}\right)_{ki} = \sum_j (U^T_{kl})_{kj} b_{ji} = \cos\varphi \cdot b_{ki} + \sin\varphi \cdot b_{li}$$

$$\left(\tilde{A}\right)_{li} = \sum_j (U^T_{kl})_{lj} b_{ji} = (-\sin\varphi) \cdot b_{ki} + \cos\varphi \cdot b_{li}$$

Итак:

$$\left(\tilde{A}\right)_{kl} = \cos\varphi\cdot b_{kl} + \sin\varphi\cdot b_{ll} = \cos\varphi\,(a_{kk}\,(-\sin\varphi) + a_{kl}\cos\varphi) + \sin\varphi\,(a_{lk}\,(-\sin\varphi) + a_{ll}\cos\varphi) =$$

$$= a_{kl}\cos 2\varphi - (a_{kk} - a_{ll})\,\frac{1}{2}\sin 2\varphi = 0 \quad \Leftrightarrow \quad \operatorname{tg} 2\varphi = \frac{2a_{kl}}{a_{kk} - a_{ll}};^{*1)} \quad |\varphi| < \pi/4, \qquad (9)$$

элементарное вращение $U_{kl}(\varphi)$ определено.

2.3 Инвариантность сферической нормы матрицы при элементарном вращении

Рассмотрим сферическую норму матрицы A

$$S = \|A\|_E^2 = \sum_{i,j} |a_{ij}|^2 = \sum_{i,j} a_{ij}^2$$

(A — вещественная матрица).

Лемма. *Величина S не изменяется при вращении $U_{kl}(\varphi)$.*

Действительно:

1) Преобразование $B = AU$ изменяет лишь два столбца k-ый и l-ый в матрице A, причём $b_{ik}^2 + b_{il}^2 = a_{ik}^2 + a_{il}^2$ и в S нет изменения;

2) Преобразование $\tilde{A} = U^T B$ изменяет лишь две строки в матрице B, причём $\tilde{a}_{ki}^2 + \tilde{a}_{li}^2 = b_{ki}^2 + b_{li}^2$ и в S нет изменений.

Таким образом

$$\left\|\tilde{A}\right\|_E = \left\|U^T B\right\|_E = \|B\|_E = \|AU\|_E = \|A\|_E\,,$$

что и требовалось показать ■

Теперь выделим в S внедиагональные элементы

$$S = \|A\|_E^2 = S_1 + S_2 = \sum_i a_{ii}^2 + \overbrace{\sum_{\substack{i,j \\ i\neq j}} a_{ij}^2}^{\text{внедиаг. эл-ты}}\,.$$

При повороте $U_{kl}(\varphi)$ (9) часть $S_2 \downarrow$ — убывает, следовательно часть $S_1 \uparrow$ — растет.

*1) Если разность $a_{kk} - a_{ll}$ в (9) равена нулю, то $\cos 2\varphi = 0$ и $\varphi = \pi/4$.

Нужно подобрать такую последовательность вращений, чтобы $S_2 \to 0$ и при этом $\tilde{A}$ станет диагональной матрицей.

Выгодно уничтожать при очередном вращении наибольший по модулю внедиагональный элемент.

Для уменьшения объема вычислений поступают так:

1) Составим суммы строк (полустрок) и найдем строку с наибольшей суммой

$$S_i = \sum_{\substack{i,j \\ i \neq j}} a_{ij}^2 \quad \Rightarrow \quad S_{i_{\max}}.$$

2) В $i_{\max}$-строке найдем наибольший по модулю элемент

$$|a_{i_{\max}, j_{\max}}|.$$

3) Его и будем исключать 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

$$S_2^{(N)} \approx \left(1 - \frac{2}{n\,(n-1)}\right)^N S_2 \approx e^{-\frac{2}{n^2} N}, \quad S_2 \to 0, \quad N \to \infty.$$

Замечания:

Процесс Якоби с выбором оптимального элемента сходится к диагональной матрице Λ.

Матрица вращений Якоби на N-ой итерации дается произведением

$$U = \prod_{i=1}^{N} U_{k_i, l_i}.$$

Её столбцами являются приближения координат собственных векторов матрицы A.

Имеет место оценка числа арифметических действий для нахождения *всех* собственных значений матрицы A — $O\left(30n^2\right)$.

ГЛАВА VI

МЕТОДЫ ОПТИМИЗАЦИИ

§1. Постановка задачи оптимизации. Необходимые и достаточные условия экстремума

Говоря о задачах оптимизации выделяют несколько общих моментов:

- Определяют некоторую "скалярную" (что важно для нас) меру *качества* — *целевую функцию* "Φ".
- Определяют набор *независимых* переменных и формулируются условия, которые характеризуют их приемлемые значения (размерность задачи и её ограничения).
- Решение оптимизационной задачи — это приемлемый набор значений переменных, которому отвечает *оптимальное* значение целевой функции.

Под *оптимальностью* (в нашем рассмотрении) обычно понимают *минимальность* целевой функции.

Пусть $x \in \mathcal{M}$ — элемент метрического пространства $\mathcal{M}$ и с помощью ограничений выделено множество $\mathcal{X} \subseteq \mathcal{M}$.

Говорят, что целевая функция $\Phi(x)$ имеет локальный минимум на элементе $x^* \in \mathcal{X}$, т. е. $x^* \in loc\min\limits_{\mathcal{X}} \Phi(x)$, если существует некоторая конечная ϵ-окрестность точки x^* — шар $K_\epsilon(x^*)$, такая, что

$$\Phi(x^*) < \Phi(x), \quad \forall x \in K_\epsilon(x^*), \quad 0 < \rho(x, x^*) < \varepsilon.^{*1)} \tag{1}$$

У функции $\Phi(x)$ может быть несколько локальных минимумов — множество $loc\min\limits_{\mathcal{X}} \Phi(x)$. Если же в этом множестве существует точка $x^* \in loc\min\limits_{X} \Phi$, в которой достигается *наименьшее* значение функции

$$\Phi(x^*) = \inf_{\mathcal{X}} \Phi(x), \tag{2}$$

то говорят о достижении в т. x^* *абсолютного* минимума[*2)].

[*1)] В случае (1) говорят о *строгом минимуме* (в смысле неравенства), тогда как $\Phi(x^*) \le \Phi(x)$ при $\rho(x, x^*) < \epsilon$ говорят о нестрогом минимуме.

[*2)] Мы ограничимся именно рассмотрением такого случая, когда глобальный экстремум функции совпадает с одним из её локальных минимумов.

Относительно целевой функции $\Phi(x)$ естественно требовать её непрерывности, хотя и не всегда; а относительно множества $\mathcal{X}$ — *компактности* и *замкнутости* этого множества. Напомним:
Множество $\mathcal{X}$ — компактно, если из каждого его бесконечного и ограниченного подмножества можно выделить сходящуюся последовательность точек.
Множество $\mathcal{X}$ — замкнуто, если предел любой сходящейся последовательности точек $\{x_n\}$ из $\mathcal{X}$ принадлежит $\mathcal{X}$.

В частности при $\mathcal{X} = \mathcal{M}$ само $\mathcal{M}$ должно быть *банаховым* пространством.

Мы ограничимся рассмотрением задачи (1) о *локальном* экстремуме. Задача (2) решается выбором наименьшего из соответствующих локальных минимумов.

Второе существенное ограничение — это рассмотрение задачи минимизации без ограничений, т. е. $\mathcal{X} = \mathcal{M}$:

1) $\mathcal{X} = R^1$ — задача минимизации функции одного переменного;

2) $\mathcal{X} = R^n$ — задача минимизации функциии n переменных;

3) $\mathcal{X}$ — гильбертово пространство и задача о минимизации функционала (скалярная целевая функция).

С решением задачи (1) в предположении соответсвующей гладкости целевой функции $\Phi(x)$ связывают *необходимое* условие экстремума (Эйлера)

$$\left.\frac{\delta\Phi}{\delta x}\right|_{x^*} = 0. \tag{3}$$

Для случая одного переменного это условие приводит к одному нелинейному уравнению

$$\Phi'(x) = 0.$$

В случае n-мерной задачи мы получаем систему нелиненых уравнений

$$\frac{\partial\Phi}{\partial x_k}(x_1, ..., x_n) = 0.$$

В случае задачи минимизации функционала $\Phi(x)$ уравнение (3), как правило, дифференциальное или интегро-дифференциальное уравнение. Например, для функционала

$$\Phi(x) = \int_a^b F(t, x(t), \dot{x}(t)dt)$$

получаем

$$\delta\Phi = 0 \quad \Leftrightarrow \quad \begin{cases} \frac{d}{dt}\left(\frac{\partial F}{\partial \dot{x}}\right) - \frac{\partial F}{\partial x} = 0 & \text{уравнение} \\ + \text{краевые условия} & \text{Эйлера-Лагранжа} \end{cases}$$

Численное решение задачи (3) — отдельная самостоятельная проблема (частично нами решенная). Как правило здесь используются итерационные методы, обладающие своими достоинствами и недостатками. Нас же будут интересовать в основном методы безусловной минимизации (1), не связанные прямо с решением необходимого условия (3).

§2. Минимум функции одного переменного

2.1 Постановка задачи одномерной минимизации

Рассмотрим задачу безусловной минимизации функции одного переменного:
Требуется найти т. $x^* \in R$ *такую, что*

$$\Phi(x^*) = \min \Phi(x) \quad \Leftrightarrow \quad x^* \in loc \min_{x \in R} \Phi(x). \tag{4}$$

Если функция $\Phi(x) \in C^{(2)}(R)$ дважды непрерывно дифференцируема, то известны необходимое и достаточное условия минимума:

необходимое условие экстремума	достаточное условие экстремума
$\Phi'(x^*) = 0$	$\Phi'(x^*) = 0$
$\Phi''(x^*) \geq 0$	$\Phi''(x^*) > 0$

(5)

(Взятые по отдельности — это соответствующие условия оптимальности точки x^* первого и второго порядков как необходимые, так и достаточные).

В таком случае, при нахождении в достаточно малой окрестности точки x^*, разложение целевой функции в ряд Тейлора с центром в точке x^* имеет вид

$$\Phi(x^* + h) = \Phi(x^*) + \underbrace{\Phi'(x^*)\,h}_{\equiv 0 \text{ в силу } (5)} + \frac{1}{2!}\Phi''(x^*)\,h^2 + o(h^2).$$

Мы говорим о *невырожденности минимума* в точке x^*, если $\Phi''(x^*) \neq 0$, тем самым, согласно (5), $\Phi''(x^*) > 0$. В дальнейшем будем предполагать это условие выполненным.

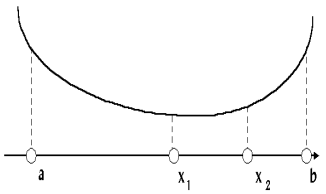
Подчеркнем еще раз, что мы пытаемся рассмотреть способы минимизации непосредственно задачи (4), а не решение задачи (5) из необходимого условия экстремума. Хотя, конечно, это тесно связаные проблемы.

2.2 Методы минимизации нулевого порядка

Под методами минимизации *нулевого порядка* подразумевают группу методов не использующих явно производные целевой функции.

Предположим, что точки a и b определяют, возможно и достаточно грубо, интервал, где расположено значение точки минимума x^* задачи (4). Если считать, что внутри этого интервала функция $\Phi(x^*)$ *унимодальна*, т. е. имеет единственный минимум, то одна из возможностей построения последовательности стягивающихся отрезков $x^* \in [x_{k-1}, x_k]$, локализующих x^* возможна на основании:
а)*метода дихотомии* — разобрать самостоятельно!
б)*метода "золотого сечения"*. Пусть на $[a, b]$ даны две внутренние точки x_1 и x_2. Сравним значение целевой функции $\Phi(x^*)$ в точках $\{a; x_1; x_2; b\}$ и выберем из них наименьшее. Пусть это $\Phi(x_1)$.

Тогда минимум функции $\Phi(x)$ — точка x^* — расположен на одном из соседних с точкой x_1 отрезков и отрезок $(x_2, b]$ можно из рассмотрения удалить.

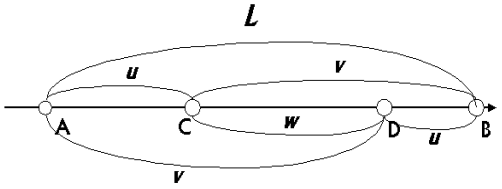


Теперь, на отрезке $[a, x_2]$, одна точка (x_1) — уже есть. Добавим следующую — x_3 и повторим отбор.

Как выгодно размещать точки? В методе *"золотого сечения"* поступают так: Напомним,

Точка C делит отрезок AB по правилу "золотого сечения", если отношение длины всего отрезка к его большей доле такое же, как и отношение длины большей доли к меньшей его доле.

$$\begin{cases} \frac{L}{v} = \frac{v}{u} \\ u + v = L. \end{cases}$$



Обозначив отношение длин отрезков через τ, найдем:

$$\begin{cases} \frac{AB}{CB} = \frac{L}{v} = \frac{v}{u} = \tau > 1 \\ u + v = L \end{cases} \quad \Leftrightarrow \quad \tau^2 - \tau - 1 = 0 \quad \Leftrightarrow \quad \tau = \frac{1+\sqrt{5}}{2} \approx 1,62. \qquad (6)$$

Если далее на большем отрезке отложить меньший отрезок $u = BD$ от точки D (или от точки C, то точка D делит BC по правилу золотого сечения, ибо $\frac{CB}{BD} = \frac{v}{u} = \tau$, причем BD — "большая" доля. Заметим, что $AD = CB$ всегда.

Итак, пусть длины L_k частичных отрезков, локализующих экстремум x^*, строились по правилу "золотого сечения". Это позволяет написать

$$\frac{L}{L_1} = \frac{L_1}{L_2} = ... = \frac{L_{k-1}}{L_k} = \frac{L_k}{L_{k+1}} = ... = \tau = \frac{1+\sqrt{5}}{2}$$

$$L_{k-1} = L_k + L_{k+1}.$$

Перемножая, найдем

$$\frac{L}{L_1} \cdot \frac{L_1}{L_2} \cdot ... \cdot \frac{L_{k-1}}{L_k} = \tau^k \quad \Leftrightarrow \quad L_k = \frac{L}{\tau^k}. \qquad (7)$$

Таким образом последовательность длин частичных отрезков $L_k \to 0$, при $k \to \infty$. Мы получили, что итерационная процедура (6-7) — процедура первой степени точности. Сходимость метода "золотого сечения" линейная (со скоростью геометрической прогрессии).

Замечание. Формулы (6-7) дают достаточно медленную сходимость. Много раз вычисляется целевая функция $\Phi(x)$, но мало используется информация о самих значениях функции $\Phi(x)$ на предыдущих шагах метода, только сравнение ">" или "<" при выборе очередной точки последовательности $\{L_k\}$.

2.3 Методы более высокого порядка (метод парабол)

Использование информации о значениях $\Phi(x)$ или её производных позволяет аппроксимировать $\Phi(x)$ многочленом в окрестности точки x_k, при этом

а) если $\Phi(x) \in C^{(2)}$, то из формулы Тейлора

$$\Phi(x) = \underbrace{\Phi(x_k) + \Phi'(x_k)(x - x_k) + \frac{1}{2!}\Phi''(x_k)(x - x_k)^2}_{\Psi(x)} + o\left((x - x_k)^2\right)$$

В качестве x_{k+1} берется точка экстремума квадратичной функции $\Psi(x)$ т. е. $\Psi'(x_{k+1}) = 0$. Получим

$$\Phi'(x_k) + (x - x_k)\Phi''(x_k) \quad \Leftrightarrow \quad x_{k+1} = x_k - \frac{\Phi'(x_k)}{\Phi''(x_k)}. \tag{8}$$

Мы получили итерационный процесс, совпадающий с методом Ньютона поиска корня уравнения $\Phi'(x) = 0$. Поэтому (8) обеспечивает не хуже, чем "квадратичную" сходимость в достаточно малой окрестности невырожденного экстремума x^*. Но на каждой итерации необходимо вычислять производные $\Phi'(x_k)$ и $\Phi''(x_k)$ целевой функции.

б) Если не прибегать к вычислению $\Phi'(x_k)$ и $\Phi''(x_k)$, то по трем точкам x_{k-2}, x_{k-1}, x_k и соответствующим значениям$\Phi(x_{k-2})$, $\Phi(x_{k-1})$ и $\Phi(x_k)$ можно построить интерполяционный многочлен Ньютона (парабола)

$$N_2(x) = \Phi(x_k) + (x - x_k)\Phi(x_k, x_{k-1}) + (x - x_k)(x - x_{k-1})\Phi(x_k, x_{k-1}, x_{k-2})$$

и вычислить x_{k+1} как точку экстремума $N_2(x)$, т.е. координату "вершины" параболы.

$$N_2'(x) = 0 \Leftrightarrow \Phi(x_k, x_{k-1}) + (2x - (x_k + x_{k-1}))\Phi(x_k, x_{k-1}, x_{k-2}) = 0$$

$$x_{k+1} = \frac{x_k + x_{k+1}}{2} - \frac{\Phi(x_k, x_{k-1})}{\Phi(x_k, x_{k-1}, x_{k-2})} = \left\{ \begin{array}{c} \text{получить} \\ \text{самостоятельно} \\ \text{окончательную} \\ \text{расчётную} \\ \text{формулу} \end{array} \right\}. \tag{9}$$

Замечания:

В обоих случаях обязательна проверка условия $\Phi(x_{k+1}) < \Phi(x_k)$ для исходной целевой функции.

Сходимость метода парабол (9) выше чем линейная, но не квадратичная:

$$|x_{k+1} - x^*| \le C\,|x_k - x^*|^{\sim 4/3}, \quad \lim_{k\to\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|} = 0.$$

§3. Минимум функции многих переменных

3.1 Постановка задачи. Необходимые и достаточные условия экстремума

1) Если целевая функция $\Phi(x) \equiv \Phi(x_1, \ldots, x_n)$, $x \in R^n$, то минимизация $\Phi(x)$ приводит к задаче:

$$x^* \in loc\min_{R^n} \Phi(x); \qquad \Phi(x^*) = \min_{K_\varepsilon(x^*)} \Phi(x). \tag{10}$$

Введем в рассмотрение градиент и гессиан функции Φ:

$$\vec{g}(x) = grad\Phi = \vec{\nabla}\Phi = \left\{ \frac{\partial\Phi}{\partial x_1}, ..., \frac{\partial\Phi}{\partial x_n} \right\},$$

$$G(x) = \operatorname{hess}\Phi(x) = \left\| \frac{\partial^2\Phi}{\partial x_i \partial x_j} \right\| \quad — \quad \begin{array}{c}\text{симметричная}\\ \text{матрица}\\ \text{вторых}\\ \text{производных}\\ \Phi(x_1, ..., x_n).\end{array}$$

Тогда разложение функции $\Phi(\vec{x})$ в ряд Тейлора в окрестности точки $\vec{x}$ при $\Delta\vec{x} = h\vec{p}$; $\|\vec{p}\| = 1$; $h = \Delta x$ имеет вид:

$$\Phi(\vec{x} + h\vec{p}) = \Phi(\vec{x}) + d\Phi(\vec{x}) + \frac{1}{2}d^2\Phi(\vec{x}) + o\left(\|\delta x\|^2\right) =$$

$$= \Phi(\vec{x}) + (grad\Phi, \vec{p})\, h + \frac{h^2}{2}\underbrace{(\operatorname{hess}\Phi(\vec{x})\cdot\vec{p}, \vec{p})}_{(G\vec{p}, \vec{p})} + o\left(h^2\right).$$

Величина $(\vec{g}(\vec{x}), \vec{p}) \equiv \frac{\partial\Phi}{\partial p}$ — производная Φ в точки $\vec{x}$ по направлению $\vec{p}$; $(G\vec{p}, \vec{p}) = (\operatorname{hess}\Phi\vec{p}, \vec{p}) \equiv K(\vec{p})$ — кривизна поверхности $u = \Phi(\vec{x})$ в точке $\vec{x}$ по направлению $\vec{p}$.

2)Необходимые и достаточные условия минимума для дважды дифференцируемой функции $\Phi(x_1, ..., x_n)$. Напомним

необходимое условие экстремума	достаточное условие экстремума
$\|\vec{g}(x^*)\| = 0$	$\|\vec{g}(x^*)\| = 0$
$\operatorname{hess}\Phi(x^*) \geq 0$	$\operatorname{hess}\Phi(x^*) > 0^{*1)}$

(11)

*1)матрица $A > 0$, если $\forall x \neq 0 (Ax, x) > 0$ — положительно определенная квадоатичная форма.

3.2 Квадратичная функция аргумента $\vec{x}$

Опираясь на тейлоровское разложение естественно в качестве удобной аппроксимации гладкой функции $\Phi(x)$ в окрестности некоторой точки (в том числе и точки возможного экстремума) использовать квадратичную функцию $\Psi(\vec{x})$:

$$\Psi(\vec{x}) = \frac{1}{2}(A\vec{x}, \vec{x}) + \left(\vec{b}, \vec{x}\right) + c,$$

где A — симметричная, невырожденная матрица $A = A^T$, $detA \neq 0$.
Установим вид градиента $\vec{\nabla}\Psi$ и гессиана $G = \operatorname{hess}\Psi$ функции $\Psi(\vec{x})$:

$$\Psi(\vec{x} + h\vec{p}) = \frac{1}{2}(A(\vec{x} + h\vec{p}), \vec{x} + h\vec{p}) + \left(\vec{b}, \vec{x} + h\vec{p}\right) + c =$$

$$= \left\{\frac{1}{2}(Ax, x) + (b, x) + c\right\} + h\left\{\frac{1}{2}(A\vec{x}, \vec{p}) + \frac{1}{2}(A\vec{p}, \vec{x}) + \left(\vec{b}, \vec{p}\right)\right\} + \frac{h^2}{2}(A\vec{p}, \vec{p}) =$$

$$= \Psi(\vec{x}) + h\left(A\vec{x} + \vec{b}, \vec{p}\right) + \frac{h^2}{2}(A\vec{p}, \vec{p}) \text{ , т.е.}$$

$$grad\Psi = A\vec{x} + \vec{b}; \quad \operatorname{hess}\Psi(\vec{x}) = A \text{ — } const. \tag{12}$$

Стационарная точка для $\Psi(\vec{x})$ удовлетворяет условию:

$$grad\Psi(x^*) = 0 \quad \Leftrightarrow \quad Ax^* + b = 0 \quad \Leftrightarrow \quad Ax^* = -b \text{ — СЛАУ} \tag{13}$$

Решение системы (13) зависит от ранга матрицы A. В случае совместной системы решение может быть и неединственным.
В окрестности стационарной точки $\vec{x}^*$:

$$\Psi(\vec{x}) = \Psi(\vec{x}^* + h\vec{p}) = \Psi(\vec{x}^*) + \frac{h^2}{2}(A\vec{p}, \vec{p}).$$

И поведение квадратичной функции определяется только свойствами матрицы A. Если A — симметричная невырожденная матрица, то существует ортонормированный базис (ОНБ) из собственных векторов матрицы A. Пусть $\{\lambda_i, \vec{x}_i\}$ — собственные значения и собственные векторы матрицы A, $\{\vec{x}_i\}$ — ОНБ. Разложим направление $\vec{p}$ по базису $\{\vec{x}_i\}$ — $\vec{p} = \sum\limits_{i=1}^{n} \alpha_i \vec{x}$, тогда

$$\Psi(\vec{x}^* + h\vec{p}) = \Psi(\vec{x}^*) + \frac{h^2}{2}\left(\sum_i \alpha_i\lambda_i\vec{x}_i, \sum_i \alpha_j\vec{x}_j\right) = \Psi(\vec{x}^*) + \frac{h^2}{2}\sum_i \lambda_i\alpha_i^2. \tag{14}$$

Характер изменения $\Psi(\vec{x})$ при движении вдоль $\vec{x}_k$ полностью определяется знаком λ_k. Если $A > 0$, то все $\lambda_i > 0$ и x^* — точка минимума.

3.3 Рельеф поверхности целевой функции $\Phi(x)$. Поверхности уровня

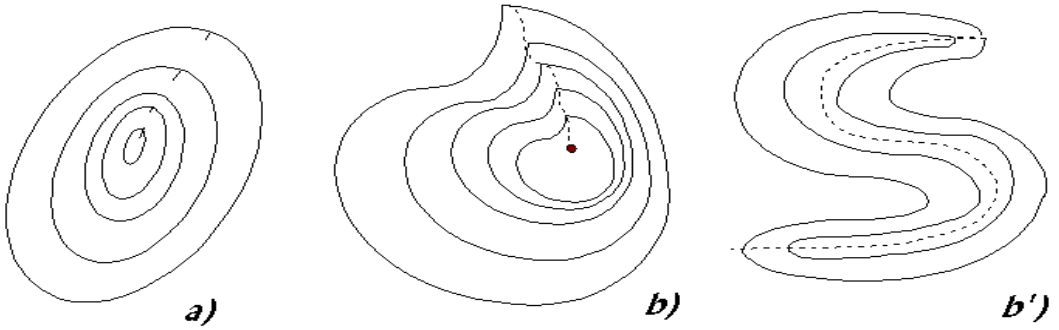
Трудности и проблемы задачи минимизации, характерные для общего случая, столь же ясно проявляются и при рассмотрении минимизации функции двух переменных $\Phi(x,y)$. Геометрию поверхности $z=\Phi(x,y)$ представляют с помощью "плоских" линий уровня

$$L_0=\{(x,y):\quad \Phi(x,y)=\Phi(x_0,y_0)=\Phi_0=const\},$$

являющихся проекциями на плоскость OXY сечения поверхности $z=\Phi(x,y)$ плоскостью $z_0=\Phi_0$.

Выделяют три основных типа рельефа поверхности.

а) *котловинный* — линии уровня похожи на концентричекские эллипсы с главными осями параллельными собственным векторам $\operatorname{hess}\Phi(x,y)$ В малой окрестности невырожденного минимума (x^*,y^*) $\operatorname{hess}\Phi(x,y)>0$ и рельеф поверхности именно *котловинный*.



б) *овражный* — если линия уровня кусочно-гладкая, то геометрическое место точек (ГМТ) излома по всем линиям уровня называют *истинным оврагом* (если угол излома направлен в сторону возрастания функции) или *истинным гребнем* (если угол излома направлен в сторону убывания функции).

Однако чаще приходится иметь дело с *разрешимыми* оврагами и гребнями (ГМТ наибольшей кривизны — рисунок b')). Например, одна из стандартных тестовых функций многомерной минимизации (функция Розенброка)

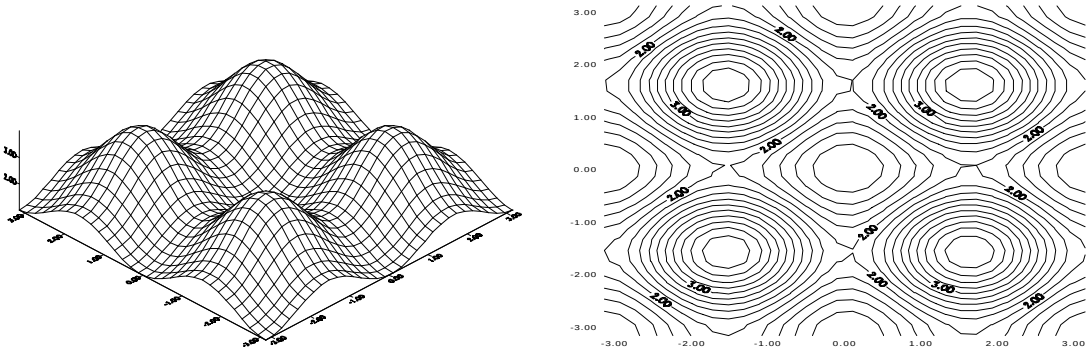
$$\Phi(x,y)=100\left(y-x^2\right)^2+(1-x)^2$$

обладает пологим серповидным ("банановидным") ущельем и имеет абсолютный минимум в точке $x^*(1,1)$.

в) *неупорядоченный тип рельефа* — характеризуется наличием многих максимумов, минимумов и седловин. Приведем в качестве примера функцию

$$\Phi(x,y)=\left(1+\sin^2 x\right)\left(1+\sin^2 y\right)$$

с достаточно неупорядоченным рельефом:



Если рассматривать дифференцируемую в каждой точке функцию $\Phi(\vec{x})$, то её производная по направлению $\vec{p}$

$$\frac{\partial \Phi}{\partial p} = (grad\Phi, \vec{p}) = \vec{g} \cdot \vec{p}$$

обладает характерными свойствами на поверхности уровня

–производная по направлению градиента — максимальна;

–вдоль линии уровня $\frac{\partial \Phi}{\partial p}$ равна нулю и градиент $\vec{g}$ перпендикулярен линии уровня в каждой точке.

3.4 Спуск по координатам

Все методы минимизации сводятся к построению траектории спуска $\{M_k\}$ вдоль которой целевая функция убывает:

$$\Phi(M_{k+1}) < \Phi(M_k)$$

(или невозрастает).

Опишем *координатный спуск*. Выберем нулевое приближение $M_0(\overset{(0)}{x_1}, ..., \overset{(0)}{x_n})$ и зафиксируем все значения координат, кроме первой. Тогда $\Phi(\vec{x})$

$$\Phi(x_1, \overset{(0)}{x_2}, ..., \overset{(0)}{x_n}) \equiv \varphi_1(x_1)$$

становится функцией одного переменного.

Используя методы минимизации функции одного переменного, найдем точку её минимума $\overset{(1)}{x_1}$ и совершим шаг из M_0 в $M_0^{(1)}(\overset{(1)}{x_1}, \overset{(0)}{x_2}, ..., \overset{(0)}{x_n})$.

На k-м шаге спуска: Из точки $M_0^{(k-1)}(\overset{(1)}{x_1}, ..., \overset{(1)}{x_{k-1}}, \overset{(0)}{x_k}, ..., \overset{(0)}{x_n})$ спускаемся по x_k минимизируя

$$\varphi_k(x_k) \equiv \Phi\left(\overset{(1)}{x_1}, ..., \overset{(1)}{x_{k-1}}, x_k, \overset{(0)}{x_{k+1}}, ..., \overset{(0)}{x_n}\right),$$

$$\overset{(1)}{x_k}: \quad \varphi(\overset{(1)}{x_k}) = \min_{x_k} \varphi_k(x_k) \tag{15}$$

в точку

$$M_0^{(k)}\left(\overset{(1)}{x_1}, ..., \overset{(1)}{x_k}, \overset{(0)}{x_{k+1}}, ..., \overset{(0)}{x_n}\right).$$

И так до тех пор, пока не выполним один цикл спуска по координатам. Последнюю точку спуска назовем $M_1 \equiv M_0(\overset{(1)}{x_1}, ..., \overset{(0)}{x_n}) \equiv M_1(\overset{(1)}{x_1}, ..., \overset{(0)}{x_n})$. Траектория $\{M_k\}$ — траектория спуска, поскольку

$$\Phi\left(M_k\right) \leq \Phi\left(M_{k-1}\right).$$

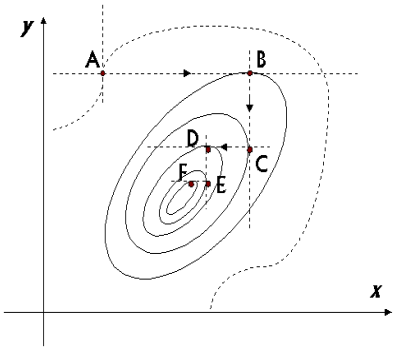
В силу ограниченности снизу значений $\Phi(x)$ значением $\Phi\left(x^*\right) \equiv \Phi^*$ (мы предполагаем, что экстремум существует), то

$$\Phi_k \geq \Phi^* \Rightarrow \lim_{k\to\infty} \Phi_k = \tilde{\Phi} \geq \Phi^* \ (!)$$

Будет ли здесь равенство, т.е. сходится ли спуск по координатам к минимуму и как быстро, зависит от функции $\Phi\left(\vec{x}\right)$ и выбранного начального приближения $\vec{x}_0$ (оно должно попасть в область влияния локального экстремума).

Рассмотрим трактовку координатного спуска на примере функции двух переменных:

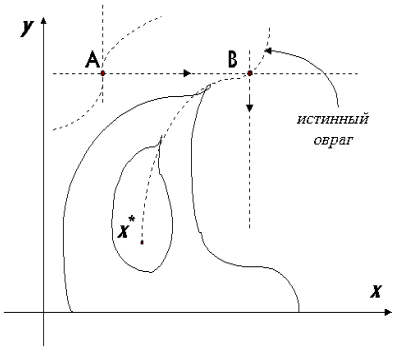
Двигаясь по прямой AB мы пересекаем линии уровня $(x, y) = const$, при этом $\Phi\left(x, y\right)$ либо возрастает, либо убывает в зависимости от направления движения. Только в одной точке B, где данная прямая касается линии уровня, функция $\Phi\left(x, y\right)$ имеет минимальное значение в данном направлении (экстремум по x или по y). Найдя такую точку, завершаем спуск по данному направлению.



Заметим, что в координатном спуске соответствующие направления взаимноортогональны.

Если в рельефе наличествует "истинный" овраг, то спуск (в данном случае первый же спуск в точку B) приводит к попаданию на "дно" оврага. А поскольку он ориентирован достаточно произвольно, то дальнейший спуск может оказаться невозможным.

Хотя минимум еще и не достигнут!



Если же $\Phi\left(\vec{x}\right)$ достаточно гладкая функция и минимум невырожден, $\operatorname{hess}\Phi\left(\vec{x}^*\right) > 0$, то в окрестности $\vec{x}^*$ рельеф котловинный и координатный спуск ведет нас к локальному минимуму при произвольном начальном приближении $\vec{x}_0$ в этой окрестности.

Рассмотрим достаточные условия сходимости координатнрого спуска на примере функции двух переменных:

Теорема 1. *Пусть D - множество уровня, ограниченное линией уровня $\Phi(x, y) = \Phi_0$, т.е.*

$$D = \{(x, y): \quad \Phi(x, y) \leq \Phi(x_0, y_0)\},$$

замкнутая ограниченная область и в D функция $\Phi\left(x, y\right)$ дважды дифференцируема, причем

$$\Phi_{xx} \geq a > 0; \ \Phi_{yy} \geq b > 0; \ |\Phi_{xy}| \leq c \ и \ ab > c^2. \ ^{*1)} \tag{15}$$

Тогда траектория координатного спуска $\{M_k\}$ (14) из произвольной точки $M_0 \in D$ сходится к локальному минимуму x^* в области D.

Доказательство.

Докажем сходимость $grad\Phi(M_k)$ на траектории спуска $\{M_k\}$. Проследим за изменением Φ_x и Φ_y на траектории спуска $\{M_k\}$. Поскольку $\Phi(x,y)$ вдоль траектории спуска не возрастает, то все точки $M_k \in D_0$. Пусть предыдущий цикл спусков закончился в точке A, тогда

$$\Phi_y(A) = 0, \quad |\Phi_x(A)| = U \neq 0.$$

Попав в точку экстремума B на прямой AB получим следующие компоненты градиента

$$|\Phi_y(B)| = V \neq 0, \quad \Phi_x(B) = 0.$$

Теперь нетрудно получить, что

$$\begin{cases} U = |\Phi_x(B) - \Phi_x(A)| = |\Phi_{xx}(\xi)| \cdot |x_B - x_A| \geq a \cdot \rho(A,B) \\ V = |\Phi_y(B) - \Phi_y(A)| = |\Phi_{xy}(\eta)| \cdot |x_B - x_A| \leq c \cdot \rho(A,B) \end{cases} \quad \Rightarrow \quad Uc \geq Va.$$

Спустившись далее по направлению BC в точку экстремума C, найдём

$$\begin{cases} V = |\Phi_y(C) - \Phi_y(B)| = |\Phi_{yy}(\xi)| \cdot |y_C - y_B| \geq b \cdot \rho(B,C) \\ W = |\Phi_x(C) - \Phi_x(B)| = |\Phi_{xy}(\eta)| \cdot |y_C - y_B| \leq c \cdot \rho(B,C) \end{cases} \quad \Rightarrow \quad Vc \geq Wb.$$

Окончательно, за один цикл спуска, получаем

$$W \leq \frac{c}{b} V \leq \frac{c^2}{ab} U = q \cdot U,$$

причём, в силу условий теоремы (15), $q < 1$.

Итак, за один цикл спусков $|\Phi_x|$ уменьшился в q раз. Аналогично, со сдвигом на $1/2$ цикла, $|\Phi_y|$ уменьшится в q раз. Выполнив n циклов координатного спуска получим, что

$$|\Phi_x|_{(n)} \leq q^n |\Phi_x|_{(0)} \quad \Longrightarrow \quad |\Phi_x| \underset{n\to\infty}{\longrightarrow} 0 \text{ и } |\Phi_y| \underset{n\to\infty}{\longrightarrow} 0.$$

Далее, в окрестности точки экстремума x^* компоненты градиента можно разложить по формуле Тейлора

$$\left.\begin{aligned} \Phi_x(M) &= \underbrace{\Phi_x(M^*)}_{\equiv 0} + \tfrac{\partial \Phi_x}{\partial x}(M^*) \cdot \Delta x + \tfrac{\partial \Phi_x}{\partial y}(M^*) \cdot \Delta y + \cdots \\ \Phi_y(M) &= \underbrace{\Phi_y(M^*)}_{\equiv 0} + \tfrac{\partial \Phi_y}{\partial x}(M^*) \cdot \Delta x + \tfrac{\partial \Phi_y}{\partial y}(M^*) \cdot \Delta y + \cdots \end{aligned} \right| \begin{aligned} &\Delta x = x - x^* \\ &\Delta y = y - y^*. \end{aligned}$$

Пренебрегая в разложении слагаемыми высших порядков, получаем линейную систему относительно приращений координат Δx и Δy. По условию теоремы (15) гессиан $G(M^*) > 0$, тем самым полученная система совместна и можно выразить Δx и Δy через линейную комбинацию компонент градиента в точке $M = M_{(n)}$. При этом $\Delta x, \Delta y \to 0$ на траектории $\{M_k\}, M_k \to M^*$.

Итак:

---

*1) $G(x,y) \geq d > 0$ в D. Используя критерий Сильвестра можно сформулировать многомерный аналог этого условия.

- Вблизи точки экстремума M^* сходимость координатного спуска и по координатам, и по градиенту *линейная* (достаточно медленная, что с практической точки зрения плохо);
- по "циклам" спусков можно делать ускорения по методу Эйткена;
- При попадании траектории спуска в разрешимый овраг расчет практически невозможен (слишком медленная сходимость при произвольной ориентации оврага относительно координатных осей). Поэтому выгоднее использовать методы, обладающие повышенным порядком точности.

3.5 Градиентные методы минимизации

В общем случае для траектории спуска $\{M_k\}$: $\Phi_{k+1} < \Phi_k$ при минимизации достаточно гладких функций можно сформулировать *достаточные* условия сходимости соответствующего метода спуска, характеризующие изменение функции Φ и её градиента на траектории $\{M_k\}$.
Пусть очередной шаг совершается вдоль направления $\overrightarrow{p_k}$ и приводит нас в точку M_{k+1}:

$$\overrightarrow{x}_{k+1} = \overrightarrow{x_k} + \overrightarrow{p_k} h_k.$$

Шаг h_k выбирается из условия минимальности $\Phi(M)$ вдоль $\overrightarrow{p_k}$

$$h_k : \ \varphi(h_k) = \min_h \varphi(h) = \min_h \Phi(\overrightarrow{x_k} + h\overrightarrow{p_k}).$$

Сформулируем достаточные условия сходимости метода спуска.
Теорема 2. *Пусть*

1) $\Phi(\overrightarrow{x})$ *– дважды дифференцируемая функция;*

2) множество уровня

$$D(\Phi(\overrightarrow{x_0})) = \{\overrightarrow{x} : \ \Phi(\vec{x}) \le \Phi(\vec{x}_0)\}$$

ограничено и замкнуто;

3) на каждой итерации

а) направление $\overrightarrow{p_k}$ *– "существенное направление спуска":*

$$\exists \beta < 0, \quad \overrightarrow{p_k}\,\overrightarrow{g_k} \le \beta < 0$$

б) $\Phi(x)$ *"существенно убывает" (т.е. выбрано соответствующее ограничение на шаг):*

$$\exists \mu_1, \quad \mu_2 : \ 0 < \mu_1 \le \mu_2 \le 1$$

$$-\mu_1 h_k \overrightarrow{g_k} \cdot \overrightarrow{p_k} \le \Phi_k - \Phi_{k+1} \le -\mu_2 h_k \underbrace{\overrightarrow{g_k} \cdot \overrightarrow{p_k}}_{\text{отиц. число}}$$

Тогда

$$\lim_{k\to 0} \|\overrightarrow{g_k}\| = 0;\ (M_k \to M^*)$$

т.е. метод спуска обладает сходимостью (как правило — линейной).

В основном соответствующие методы спуска отличаются выбором очередного направления $\overrightarrow{p_k}$ и шага h_k:

Метод "наискорейшего" спуска. Рассмотрим линейную аппроксимацию целевой функции $\Phi(\overrightarrow{x})$ в окрестности точки $\overrightarrow{x_k}$. Опираясь на формулу Тейлора:

$$\Phi(\overrightarrow{x_k} + \overrightarrow{p}) = \Phi(\overrightarrow{x_k}) + (grad\Phi(\overrightarrow{x_k}), \overrightarrow{p}) + o(\|\overrightarrow{p}\|),$$

с определенной точки зрения (локально!) естественно искать напрвление, по которому $\frac{\partial\Phi}{\partial p} \equiv \overrightarrow{g_k} \cdot \overrightarrow{p}$ наибольшее по модулю отрицательное число. Это направление в первом порядке по $\|\overrightarrow{p}\|$ обеспечивает наибольшее убывание функции Φ.

Итак, необходимо найти направление $\vec{p}$

$$\begin{cases} \min(\overrightarrow{g_k} \cdot \overrightarrow{p}) \\ \|\overrightarrow{p}\| = 1 \end{cases} \quad — \quad \begin{array}{c}\text{задача на}\\ \text{условный}\\ \text{экстремум}\\ \text{для } \vec{p}\end{array}$$

Решение полученной задачи зависит от вида рассматриваемой нормы. Если выбрать C-энергетическую норму $\|\overrightarrow{p}\|^2 = (C\overrightarrow{p}, \overrightarrow{p})$, где $C > 0$ и симметрична, тогда направление $\overrightarrow{p}$ (с точностью до нормировочной $Const$)

$$\overrightarrow{p} = -C^{-1} \cdot \overrightarrow{g_k}.^{*1)}$$

Для евклидовой нормы — $C \equiv E$ и $p = -\overrightarrow{g_k}$, что приводит нас к *методу наискорейшего спуска.*

$$\begin{cases} \overrightarrow{x}_{k+1} = \overrightarrow{x_k} - h_k\overrightarrow{g_k} \\ h_k:\ \varphi(h_k) = \min\limits_h \Phi(\overrightarrow{x_k} - h\overrightarrow{g_k}) \end{cases} \tag{16}$$

Замечания:

1) При таком выборе $\overrightarrow{p_k}$ и h_k (16) траектория спуска перпендикулярна линии уровня $\Phi(x_k)$ в точке x_k.

2) По сходимости *наискорейший спуск* не лучше, чем координатный спуск, т.е. он обладает лишь линейной сходимостью.

3) Анализ сходимости наискорейшего спуска на квдратичной функции с симметричной и положительно определенной матрицей (что характерно для гессиана в окрестности невырожденного минимума)

$$\Psi(x) = \frac{1}{2}(Ax, x) + (\overrightarrow{b}, x) + C:\ A > 0,\ A^T = A$$

*1)Показать!

дает лишь линейную сходимость. Поскольку $A > 0, A^T = A$ следовательно все собственные значения матрицы A положительны $\forall i \quad \lambda_i(A) > 0$. Сходимость метода наискорейшего спуска характеризуют величиной

$$\varkappa = \frac{\lambda_{max}(A)}{\lambda_{min}(A)} = \|A\| \cdot \|A^{-1}\| = CondA$$

$$\Psi(x_{k+1}) - \Psi(x^*) \simeq \left(\frac{\varkappa - 1}{\varkappa + 1}\right)^2 (\Psi(\overrightarrow{x_k}) - \Psi(x^*)) . \tag{17}$$

Полученная оценка скорости сходимости, например для $\varkappa = 100$ (хорошая обусловленность матрицы A) даёт $q \approx 0,96(!)$ и нужны сотни итераций для уменьшения погрешности на порядок.

Расчетные формулы наискорейшего спуска (16) в этом случае принимают вид:

$$\overrightarrow{g} = A\overrightarrow{x} + \overrightarrow{b};\ hess\Psi = A \quad \Rightarrow \quad \overrightarrow{p_k} = -\overrightarrow{g_k},$$

$$\psi(h) = \Psi(\overrightarrow{x} + h\overrightarrow{p}_k) = \Psi(\overrightarrow{x}) + h(Ax + b, \overrightarrow{p}_k) + \frac{h^2}{2}(A\vec{p}_k, \vec{p}_k),$$

$$\frac{\partial\psi}{\partial h} = 0 \Leftrightarrow h_k = \left\{ \begin{array}{c} \text{получить} \\ \text{самостоятельно} \\ \text{расчетные} \\ \text{формулы} \end{array} \right\} . \tag{18}$$

Тем не менее:

1) Необходимо бесконечное число итераций для нахождения экстремума даже в случае квадратичной функции.

2) Метод наискорейшего спуска не рекомендуется как серьезная минимизационная процедура. Дело в том, что свойство наискорейшего спуска является лишь *локальным* свойством, поэтому необходима частая смена направлений спуска и относительно малый шаг движения по каждому направлению, что и приводит в итоге к неэффективной вычислительной процедуре (например в случае разрешимого оврага).

3) Метод наискорейшего спуска невозможно адаптировать для использования информации о вторых производных $\Phi(\overrightarrow{x})$.

3.6 Методы второго порядка

Ньютоновские методы. Эта группа методов основана на более точной аппроксимации целевой функции в окрестности точки $\overrightarrow{x_k}$

$$\Phi(\overrightarrow{x_k} + \overrightarrow{p}) = \underbrace{\Phi(\overrightarrow{x_k}) + \overrightarrow{g_k} \cdot \overrightarrow{p} + \frac{1}{2}(G_k\overrightarrow{p}, \overrightarrow{p})}_{\Psi(\overrightarrow{p})} + o\left(\|\overrightarrow{p}\|^2\right) .$$

Минимизируемая функция $\Psi(\overrightarrow{p})$. Соответствующее направление и шаг берут из условия минимума $\Psi(\overrightarrow{p})$:

$$\left.\begin{array}{rcl} grad\Psi = 0 \quad \Leftrightarrow \quad G_k\overrightarrow{p} + \overrightarrow{g_k} = 0; \quad \Leftrightarrow & \underbrace{\overrightarrow{p_k} = -G_k^{-1}\cdot\overrightarrow{g_k}}_{\text{Ньютоновское направление}} & \\ & \overrightarrow{x}_{k+1} = \overrightarrow{x_k} + \overrightarrow{p_k} = x_k - G_k^{-1}\cdot\overrightarrow{g_k} & \end{array}\right\} \qquad (19)$$

- Для квадратичной целевой функции $\Psi(\overrightarrow{p})$ метод (19) решает задачу минимизации за одну(!) итерацию.
- В окрестности невырожденного экстремума имеет *квадратичную* сходимость (гессиан $G_k > 0$ и симметричен).
- Ньютоновское направление – это направление *наискорейшего* спуска в G-энергетической метрике
$$\|\overrightarrow{p}\| = \sqrt{(G\overrightarrow{p}, \overrightarrow{p})}.$$
- Существенным является то, что на каждом шаге необходимо решать систему линейных уравнений (19) для определения *ньтоновского направления* очередной итерации.
- При модификации метода Ньютона, когда гессиан фиксируется на определенное число итераций G_{k_0} — в методе Ньютона-Рафсона — существенен алгоритмический выйгрыш, но при этом обеспечена лишь линейная сходимость метода.

Метод сопряженных градиентов. Методы *координатного спуска* или *наискорейшего спуска* требовали даже для минимизации квадратичной функции бесконечного числа итераций.

Опираясь на тейлоровское разложение в окрестности невырожденного экстремума x^* выгодно строить методы спуска, которые, по крайней мере, эффективны для квадратичных функций.

Такими методами, не требующими решения СЛАУ (19) на каждом итерационном шаге для определения направления спуска, являются методы *сопряженных направлений.*

Для квадратичной функции $\Psi(\overrightarrow{x})$:

$$\Psi(x) = \frac{1}{2}(Ax, x) + (b, x) + c, \quad A > 0, \quad A^T = A$$

они позволяют не более чем за n шагов спуска получить её минимум. Напомним:

Симметричная положительноопределенная матрица $A > 0,\ A^T = A$ – позволяет ввести "A-энергетическю" норму вектора

$$\|x\|_A = \sqrt{(Ax, x)}$$

и соответствующеее скалярное произведение

$$(x, y)_A = (Ax, y) = (x, Ay).$$

Определение *Векторы, ортогональные в A-энергетическом смысле, называются сопряженнымы относительно матрицы A.*

$$x \underset{A}{\perp} y \quad \Leftrightarrow \quad (x,y)_A = (Ax,y) = (x,Ay) = 0.$$

Сопряженные векторы обладают рядом "хороших" свойств:

1) *Если* $\{x_i\}_k$ *– система сопряженных векторов и* $k \le n$*, то эта сисетма векторов — линейно независима.*

   Действительно, пусть $\overrightarrow{x_1} = \sum\limits_{i=2}^{k} \alpha_i \overrightarrow{x_i}$ – ненулевая комбинация остальных векторов. Тогда

   $$(x_1, Ax_1) = (x_1, A\sum_{i=2}^{k} \alpha_i \overrightarrow{x_i}) = \sum_{i=2}^{k} \alpha_i (x_1, Ax_i) \equiv 0$$

   но $A > 0$ и следовательно $\overrightarrow{x_1}$ нулевой вектор, что невозможно ■

2) Если число векторов в рассматриваемой системе $k = n$, то $\{x_i\}_n$ – сопряженный базис. Можно считать его сопряженным ОНБ, т.е. $(x_i, x_j)_A = \delta_{ij}$. Разложим направление $\overrightarrow{p}$ по ОНБ $\{x_i\}_n$ и рассмотрим квадратичную функцию на этом направлени

   $$\Psi(\overrightarrow{x} + \overrightarrow{p}) = \Psi(\overrightarrow{x}) + (Ax + b, \overrightarrow{p}) + \frac{1}{2}(A\overrightarrow{p}, \overrightarrow{p}) = \left| p = \sum \alpha_i \overrightarrow{x_i} \right| =$$

   $$= \Psi(\overrightarrow{x}) + (Ax + b, \sum_i \alpha_i \overrightarrow{x_i}) + \frac{1}{2}\left(A\sum_i \alpha_i \overrightarrow{x_i}, \sum_k \alpha_k \overrightarrow{x_k} \right) =$$

   $$= \underbrace{\sum_i \left\{ \frac{1}{2}\alpha_i^2 + \alpha_i (Ax + b, x_i) \right\}}_{n \text{ независимых слагаемых}} + \Psi(\overrightarrow{x}); \qquad (*)$$

   Движение по каждому из сопряженных направлений x_i изменяет только одно слагаемое в сумме $(*)$ и, тем самым, за не более, чем n шагов приводит к минимуму функции Ψ.

Существуют различные способы построения сопряженных относительно A направлений, в частности – метод *сопряженных градиентов* (метод Флетчера-Ривса) – приводит к одной из наиболее эффективных процедур многомерной численной минимизации.

Рассмотрим снова квадратичную аппроксимацию $\Psi(x)$ целевой функции $\Phi(x)$ в окрестности точки $\overrightarrow{x_k}$:

$$\Phi(\overrightarrow{x_k} + \overrightarrow{p}) = \underbrace{\Phi(x_k) + (grad\Phi(x_k), \overrightarrow{p}) + \frac{1}{2}(hess\Phi(x_k)\overrightarrow{p}, \overrightarrow{p})}_{\Psi_k(\overrightarrow{p})} + o\left(\|\overrightarrow{p}\|^2\right).$$

На каждом *цикле* итерационных шагов для построения *сопряженного базиса* будем использовать одну и ту же матрицу $G_k \equiv hess\Phi(x_k)$. При этом мы будем считать, что находимся в достаточно малой окрестности точки минимума x^*, где $G(x_k) > 0$.

В методе *сопряженных градиентов* совокупность сопряженных относительно $G \equiv G(x_k)$ направлений строится следующим образом. Опишем процедуру построения одного цикла минимизации, содержащего n шагов и точно минимизирующего $\Psi_k(\overrightarrow{p})$.

$$
\begin{array}{ll}
\text{Цикл движения} & M_k \equiv \overset{(1)}{M_k} \xrightarrow[\overrightarrow{p_1}]{} \overset{(2)}{M_k} \xrightarrow[\overrightarrow{p_2}]{} \cdots \xrightarrow[\overrightarrow{p}_{n-1}]{} \overset{(n)}{M_k} \xrightarrow[\overrightarrow{p_n}]{} M_{k+1} \\
\text{1-ый шаг:} & \overrightarrow{p_1} = -\overrightarrow{g_1};\ \overset{(2)}{x_k} = \overset{(1)}{x_k} + h_1 \overrightarrow{p_1};\ h_1 : \psi(h_1) = \min\limits_h \Psi_k \left(\overset{(1)}{x_k} + h \overrightarrow{p_1} \right); \\
\text{2-ой шаг:} & \overrightarrow{p_2} = -\overrightarrow{g_2} + \alpha_1 \overrightarrow{p_1};\ \alpha_1 = \frac{(g_2, g_2)}{(g_1, g_1)};\ \overrightarrow{p_2} \perp \overrightarrow{p_1} \text{ отн-но } G_k
\end{array}
\tag{20}
$$

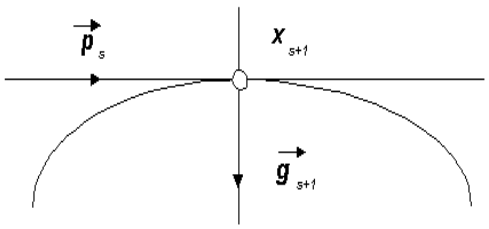
Пусть $\overrightarrow{p_1}, \ldots, \overrightarrow{p_s}$ G_k-сопряженная система векторов

$$
\left.
\begin{array}{ll}
S+1\text{-ый шаг:} & \overrightarrow{p}_{S+1} = -\overrightarrow{g}_{S+1} + \alpha_S \overrightarrow{pS} \text{ (все остальные} \alpha_i = 0) \\
& \alpha_S = \frac{g^2_{S+1}}{g^2_S};\ \text{(из сообр. } (\overrightarrow{p}_{S+1}, \overrightarrow{pS})_{G_k} = 0) \\
& \vec{x}_{S+2} = \overrightarrow{x}_{S+1} + h_{S+1} \cdot \overrightarrow{p}_{S+1} \\
& h_{S+1} :\ \psi(h_{S+1}) = \min\limits_h \Psi_k \left(\overrightarrow{x}_{S+1} + h \overrightarrow{p}_{S+1} \right)
\end{array}
\right\}
$$

Покажем, что (20) определяет систему сопряженных относительно G_k векторов движения $\{\vec{p}_s\}_n$.

а) Проверить самостоятельно 2-й шаг;

б) 1: $\overrightarrow{g}_{S+1}$ ортогонально всем предыдущим $\overrightarrow{p_j}$ при $j \le s$, ибо спускаясь на предыдущем, S-ом шаге, мы пришли в точку $\vec{x}_{S+1} = \vec{x}_S + h_S \overrightarrow{pS}$ вдоль направления $\overrightarrow{pS}$.



Но эта точка — $\overrightarrow{x}_{S+1}$ — точка "минимума", т.е. $\overrightarrow{g}_{S+1} \perp \overrightarrow{pS}$, $(\overrightarrow{g}_{S+1}, \overrightarrow{pS}) = 0$. Если проследить "вглубь" траектории, то

$$
\vec{x}_{S+1} = \vec{x}_S + h_S \vec{p}_S = \vec{x}_{S-1} + h_{S-1} \vec{p}_{S-1} + h_S \vec{p}_S = \cdots = \vec{x}_{j+1} + \sum_{j+1}^{S} h_i \overrightarrow{p_i}, \quad 1 \le j \le S-1.
$$

Тогда

$$
G x_{S+1} = G x_{j+1} + \sum_{j+1}^{S} h_i G \overrightarrow{p_i}.
$$

Добавим слева и справа по $\overrightarrow{b} \equiv \overrightarrow{g}(M_k)$, и учтём, что $G_k \equiv G$; $Gx + b \equiv \overrightarrow{g}(x)$. Таким образом

$$\overrightarrow{g}_{S+1} = \overrightarrow{g}_{j+1} + \sum_{j+1}^{S} h_i G p_i.$$

Тогда

$$\overrightarrow{g}_{S+1} \cdot \overrightarrow{p_j} = \underbrace{\overrightarrow{g}_{j+1} \cdot \overrightarrow{p_j}}_{=0 \text{ для этого шага}} + \sum_{j+1}^{S} h_i \cdot \underbrace{(Gp_i, p_j)}_{=0 \text{ в силу индукции}} \quad \Rightarrow$$

$$\Rightarrow \quad (\vec{g}_{S+1}, \vec{p}_j) = 0,\ 1 \le j \le S-1;\ S.$$

2: Покажем, что вектор $\vec{g}_{S+1}$ ортогонален всем градиентам $\vec{g}_j, j = \overline{1,S}$. Имеем

$$\overrightarrow{p_j} = -g_j + \alpha_{j-1}\overrightarrow{p}_{j-1} \quad \Leftrightarrow \quad \underbrace{\overrightarrow{g}_{S+1} \cdot \overrightarrow{p_j}}_{=0} = -(g_{S+1}, g_j) + \underbrace{\alpha_{j-1} \cdot (\overrightarrow{g}_{S+1}, \overrightarrow{p}_{j-1})}_{=0}$$

т.о.

$$(\overrightarrow{g}_{S+1}, \overrightarrow{g_j}) = 0,\ j = \overline{1,S}.$$

3: Рассмотрим очередное направление:

$$\overrightarrow{p}_{S+1} = -\overrightarrow{g}_{S+1} + \alpha_S \overrightarrow{p_S}; \quad \alpha_S = \frac{g_{S+1}^2}{g_S^2}$$

и покажем, что $\overrightarrow{p}_{S+1}$ сопряжено всем $\overrightarrow{p_j}$, $j \le S$. Оно сопряжено, по крайней мере, со всеми $\overrightarrow{p}$ до предыдущего, т.е. $(\overrightarrow{p}_{S+1}, \overrightarrow{p_j})_{G_k} = 0$, $j = \overline{1, S-1}$. Действительно

$$(\overrightarrow{p}_{S+1}, G\overrightarrow{p_j}) = \left(-\overrightarrow{g}_{S+1} + \alpha_S \overbrace{\overrightarrow{p_S}, G\overrightarrow{p_j}}^{\text{сопряжены}}\right) = -\left(g_{S+1}, G\frac{x_{j+1} - x_j}{h_j}\right) =$$

$$= -\left(g_{S+1}, \frac{(Gx_{j+1} + b_k) - (Gx_j + b_k)}{h_j}\right) = -\left(g_{S+1}, \frac{g_{j+1} - g_j}{h_j}\right) \equiv 0, \text{ ибо } j \le S-1.$$

Предыдущее направление:

$$(\overrightarrow{p}_{S+1}, G\overrightarrow{p_S}) = -(g_{S+1}, Gp_S) + \alpha(p_S, Gp_S) =$$

$$= -\left(g_{S+1}, G\frac{x_{S+1} - x_S}{h_S}\right) + \alpha_S \left(-g_S + \alpha_{S-1}\overrightarrow{p}_{S-1}, \frac{g_{S+1} - g_S}{h_S}\right) =$$

$$= -\frac{g_{S+1}^2}{h_S} + \frac{g_{S+1}^2}{\cancel{g_S^2}} \frac{\cancel{g_S^2}}{h_S} = 0 \ \blacksquare$$

Метод Флетчера-Ривса обладает квадратичной сходимость в достаточно малой корестности точки $\vec{x}^*$. Рестарт в точке M_k осуществляется по антиградиенту $(-\overrightarrow{g_k})$.

Это один из наиболее эффективных методов численной минимизации функций многих переменных.

§4. Задача минимизации функционала

4.1 Постановка задачи

Если любому $y(x) \in Y$ поставлено в соответствие число , то говорят, что на множестве функций Y задан функционал $\Phi[y(x)]$.

В задаче минимизации функционала требуется: *найти* $y^*(x) \in Y$, *на которой функционал достигает своей точной нижней грани (абсолютный экстремум)*

$$y^* : \quad \Phi^* \equiv \Phi[y^*(x)] = \inf_Y \Phi[y(x)]. \tag{21}$$

В такой постановке (21) называется задачей *минимизации по аргументу*, в отличие от

$$\Phi^* \equiv \inf_Y \Phi[y(x)] \tag{21'}$$

задачи *минимизации значений* функционала.

Не всякий функционал и не на всяком множестве имеет минимум. Скажем, если функционал неограничен снизу на заданном множестве, или соответствующее множество некомпактно в себе, или если функционал разрывен и т.д. Мы не будем исследовать постановку задачи (21), а будем предполагать, что (21) поставлена корректно, то есть её решение $y^*(x)$ на Y существует, единственно и устойчиво относительно малых возмущений входных данных.

Постановка задачи (21) возникает, как правило, когда сама модель сформулирована соответствующим образом, например рассматривается функционал "действия" (или нечто похожее)

$$\Phi[y(x)] = \int_a^b F(x, y, y', ..., y^{(p)})\, dx. \tag{*}$$

Обычно к задаче (21) приводит использование вариационных методов решения "операторного" уравнения $A[y(x)] = f(x)$. Рассмотренный нами *метод наименьших квадратов* даёт задачу минимизации функционала "невязки" для этого уравнения

$$\Phi[y(x)] = \| Ay - f \|^2 = \int_a^b (Ay(x) - f(x))^2 \rho(x) dx, \quad \rho(x) > 0.$$

В случае *некорректно* поставленной задачи $A[y(x)] = f(x)$ её *регуляризация* приводит к задаче минимизации *сглаживающего функционала Тихонова*

$$M_\alpha[y(x)] = \| Ay - f \|^2 + \alpha \Omega[y(x)],$$

где $\Omega[y(x)]$ — функционал со свойствами нормы (то есть с его помощью на Y вводится структура нормированного пространства и множество $\Omega[y] \le Const$ компактно в Y в введенной метрике). Тогда решение задачи минимизации $y_\alpha(x)$ при определённом способе согласования параметра регуляризации α с априорной информацией о

величине погрешности входных данных обладает сходимостью. Такое согласование в случае линейного оператора A возможно, например, по принципу *невязки*, что позволяет определить α из условия

$$\Phi(y^*_{\alpha(\delta)}(x)) = \| Ay^*_{\alpha(\delta)} - f \|^2 = \delta^2 = \| f - \tilde{f} \|^2,$$

При этом экстремаль $y^*_{\alpha(\delta)}(x)$ равномерно стремится к функции $y^*(x)$ при $\delta \to 0$ и поставленная для функционала $M_\alpha[y(x)]$ задача (21) корректна.

В курсе *вариационного исчисления* развиты методы минимизации функционалов (21), здесь же хотелось бы обсудить численную реализацию "прямых" методов минимизации.

4.2 Метод пробных функций

Общая идея минимизации (21) методом пробных функций состоит в сведении задачи (21) к задаче нахождения экстремума функции многих переменных. Пусть

$$Y_n = \{y_n(x;\vec{a}) \equiv y_n(x;a_1,...,a_n) \mid x \in X,\ a_i \in R\}$$

класс пробных функций, зависящих от параметров $\vec{a} = \{a_1,\ldots,a_n\}$. Тогда на функциях класса Y_n

$$\Phi[y_n(x;\vec{a})] \equiv F_n(a_1,...,a_n) \equiv F_n(\vec{a})$$

и соответствующая задача минимизации (21) ставится как задача минимизации функции $F_n(\vec{a})$

$$\vec{a}^* : \Phi_n \equiv F_n(\vec{a}^*) = \inf_{\vec{a}\in R_n} F_n(\vec{a}) = \inf_{y\in Y_n} \Phi[y_n(x;\vec{a})]. \tag{22}$$

После её решения следующий шаг – это осуществление для $y^*_n(x;\vec{a}^*)$ предельного перехода при $n \to \infty$. Поступают следующим образом:

1) Построим систему вложенных классов функций $Y_n \subset Y$:

$$Y_1 \subset Y_2 \subset Y_3 ... \subset Y_n \subset ... \subset Y;$$

2) На каждом Y_n решим задачу (22) и определим $y^*_n(x;\vec{a}^*) \equiv y^*_n(x)$ и Φ_n

$$\Phi_n \equiv \Phi[y^*_n(x)] = \inf_{Y_n} \Phi[y_n(x;\vec{a})]$$

В силу вложенности классов Y_n соответствующие $\{\Phi_n\}$ образуют по крайней мере невозрастающую последовательность:

$$\Phi_1 \geq \Phi_2 \geq ... \geq \Phi_n ... \geq \Phi^* = \inf_Y \Phi[y].$$

Построенная последовательность Φ_n сходится

$$\lim_{n\to\infty} \Phi_n = \bar{\Phi} \geq \Phi^*.$$

Если $\bar{\Phi} = \Phi^*$, то соответствующая последовательность $\{y^*_n\}$ называется минимизирующей для задачи (21) (сходимость самой последовательности $\{y^*_n\}$ не гарантируется).

Отдельно ставится вопрос о минимизации (21) по аргументу, то есть построении такой последовательности $\{y_n^*\}$, что

$$\lim_{n\to\infty} y_n^*(x) = y^*.$$

Сформулируем достаточные условия ответа на эти вопросы.

Теорема 1. *Если система функций $\{y_n(x;\vec{a})\}$ полна в Y, а функционал $\Phi[y(x)]$ непрерывен на Y, то $\{y_n^*(x;\vec{a}^*)\}$ — минимизирующая последовательность:*

$$\Phi_n \to \Phi^*.$$

Теорема 2. *Если выполнены условия Теоремы 1 и функционал $\Phi[y]$ в окрестности точки y^* удовлетворяет условию*

$$\Phi[y] - \Phi[y^*] \geq \alpha \parallel y - y^* \parallel^{\beta}; \quad \alpha, \beta > 0,$$

то

$$y_n^*(x;\vec{a}^*) \to y^*,$$

(сходимость в той норме, в которой система $\{y_n(x;\vec{a})\}$ полна).

Поясним условия сформулированных теорем. *Функционал $\Phi[y(x)]$ непрерывен в точке $y_0(x)$,если $\forall\varepsilon > 0$, $\exists\delta > 0$ такие, что*

$$\parallel y(x) - y_0(x) \parallel < \delta \quad \Longrightarrow \quad |\Phi[y(x)] - \Phi[y_0(x)]| < \varepsilon.$$

Свойство непрерывности функционала зависит и от самого функционала, и от соответствующей нормы. Например для функционала

$$\Phi[y] = \int_a^b F(x, y, y', ..., y^{(p)})dx,$$

где F — непрерывная функция по своим аргументам, имеем:
1) В норме $C^{(p)}[a,b]$, $\quad p > 0$

$$\parallel y \parallel_{C^{(p)}[a,b]} = \max_{[a,b]}\{|y|, |y'|, ..., |y^{(p)}|\}$$

$\Phi[y]$ очевидно нерпрерывен.
2) В норме $C[a,b]$

$$\parallel y \parallel_{C^{(p)}[a,b]} = \max_{[a,b]} |y(x)|$$

очевидно нет [*1)].

Бесконечную систему $\{y_n(x;\vec{a})\}$ называют полной в Y, если $\forall y(x) \in Y$ и $\forall\varepsilon > 0$ $\exists N$: $\quad \forall n > N$ (в любом классе $Y_{n>N}$) существует $y_n(x;\vec{a}) \in Y$: $\parallel y - y_n(x;\vec{a}) \parallel < \varepsilon$.

Свойство полноты сисемы функций также зависит и от самой системы функции $\{y_n(x;\vec{a})\}$ и от выбранной нормы. Если же система функций $\{y_n(x;\vec{a})\}$ полна в Y и $\Phi[y]$ — непрерывен на Y, то для y^* и $\varepsilon > 0$ найдётся $\delta > 0$ такое, что если

$$\parallel y - y^* \parallel < \delta, \text{ то } 0 \leq \Phi[y] - \Phi[y^*] < \varepsilon.$$

[*1)]Производные у "близких" в норме $C[a,b]$ функций могут сильно отличаться.

В силу полноты системы функций $\{y_n(x;\vec{a})\}$ — по указанному $\delta > 0 \Rightarrow \exists N$ такое, что $\forall n > N \; \exists \quad \tilde{y}_n(x;\vec{a})$

$$\| \tilde{y}_n(x;\vec{a}) - y^* \| < \delta$$

и

$$0 \leq \Phi[\tilde{y}] - \Phi[y^*] < \varepsilon.$$

Это неравенство верно и для $\inf_{Y_n} \Phi$.
Итак $\Phi_n - \Phi^* < \epsilon, \forall \epsilon > 0$ при $n > N$, то есть $\Phi_n \to \Phi^*$ ■
В силу условий Теоремы 2 $\exists \; \alpha, \; \beta$ такие, что имеет место

$$\Phi_n - \Phi^* \geq \alpha \| y_n^* - y^* \|^{\beta},$$

то есть

$$\forall \varepsilon > 0 \quad \Rightarrow \quad \| y_n^* - y^* \| \leq (\frac{\varepsilon}{\alpha})^{\frac{1}{\beta}} = \varepsilon_1.$$

Таким образом

$$y_n^* \to y^* \quad \text{при } n \to \infty \quad \blacksquare$$

Замечания:

Обычно вложенные классы функций $\{Y_n\}$ строят как линейные оболочки $L(y_1, ..., y_n)$ совокупности "базисных" функций $\{y_i(x)\}$.

Приведем пример полной в $C^{(p)}[a,b]$ системы функций. Рассмотрим систему сплайнов на $[a,b]$ порядка не ниже p на сгущающихся сетках ω_n — $\{s^n_{\geq(p)}(x)\}$. В норме $C^{(p)}$ эта система функций полна.

В качестве применения метода пробных функций рассмотрим:

4.3 Метод Ритца решения уравнения $Ay = f$

Пусть оператор A рассматриваемой задачи положителен и симметричен

$$Ay = f \tag{23}$$

($A^* = A, A > 0$). Покажем, что задача (23) эквивалентна задаче минимизации квадратичного функционала

$$Ay = f \quad \Leftrightarrow \quad \Phi[y] = \| y \|^2_A - 2(y,f) = (y, Ay) - 2(y,f) = \min. \tag{24}$$

Действительно $\forall y$ представим $y = y^* + \lambda\delta y$, где y^*–пока произвольно. Тогда

$$\Phi[y^* + \lambda\delta y] = \Phi[y^*] + 2\lambda(\delta y, Ay^* - f) + \lambda^2(\delta y, A\delta y).$$

Если y^* решение задачи (23), то

$$2\lambda(\delta y, Ay^* - f) = 0,$$

но $(\delta y, A\delta y) \geq 0$ по крайней мере, т.о.

$$\inf_{\delta y} \Phi[y^* + \lambda\delta y] = \Phi[y^*]$$

Это и означает, что y^* решение задачи (24).
Обратно, если y^* — решение задачи (24), то вариация функционала

$$\delta\Phi[y^*] = 0 \quad \Leftrightarrow \quad \frac{d}{d\lambda}\Phi[y^* + \lambda\delta y]|_{\lambda=0} = 2(\delta y, Ay^* - f) = 0, \quad \forall \delta y.$$

В частности и для $\delta y = Ay^* - f$, то есть $\| Ay^* - f \| = 0$, или $Ay^* = f$. Таким образом y^* есть решение задачи (23) ■

Выберем пробные функции $\{y_n(x;\vec{a})\}$ в виде

$$y_n(x;\vec{a}) = \sum_{k=1}^{n} a_k \varphi_k(x), \tag{25}$$

где $\{\varphi_k(x)\}$ –полная в Y система функций. Тогда функционал (24) принимает вид:

$$\Phi[y_n(x;\vec{a})] \equiv F_n(\vec{a}) = (y, Ay) - 2(y, f)|_{y_n(x,\vec{a})} =$$

$$= \sum_k \sum_m a_k a_m (\varphi_k, A\varphi_m) - 2\sum_k a_k(\varphi_k, f) \Rightarrow \min_{\{a_k\}} L(a_k). \tag{$*$}$$

Задача (*) — задача о минимуме квадратичной функции на R_n. Используем уравнение Эйлера $\partial L/\partial a_p = 0$. Получим

$$\sum_{m=1}^{n} a_m(\varphi_p, A\varphi_m) = (\varphi_p, f), \quad p = \overline{1,n}. \tag{26}$$

СЛАУ (26) позволяет определить параметры $\{a_p\}_m$ и решает задачу (*).
Замечания:
Мы оставим в стороне дискуссию о выборе "базисных" функций . Заметим, что если $A : F \Rightarrow F$, причём F полно в энергетической норме $A : (u,v)_A = (u, Av)$, то решение задачи (23) существует и $y_n(x;\vec{a}) \to y^*$ в A-энергетической норме.

ГЛАВА VII

ОБЫКНОВЕННЫЕ ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ

§1. Задача Коши

1.1 Постановка задачи

При решении многих задач естествознания в качестве математической модели используется *задача Коши* для обыкновенных дифференциальных уравнений. Например задачи динамики системы взаимодействующих тел (в модели материальных точек), задачи химической кинетики, электрических цепей. Ряд важных уравнений в частных производных в случаях, допускающих разделение переменных, приводит к задачам для обыкновенных дифференциальных уравнений — это, как правило, краевые задачи (задачи о собственных колебаниях упругих балок и пластин, определения спектра собственных значений энергии частицы в сферически-симметричных полях и многие другие).

Мы ограничимся рассмотрением лишь задачи Коши. Полученная в общем случае задача для ОДУ (обыкновенных дифференциальных уравнений) с помощью замены переменных сводится к *нормальной системе* дифференциальных уравнений. *Задача Коши* для последней формулируется так:
Определить дифференцируемую функцию $u(x)$, для которой

$$\frac{d\,u}{d\,x} = f(x,u) \tag{1}$$

и выполнено начальное условие

$$u(x_0) = u_0. \tag{2}$$

Здесь x_0, y_0 — заданные величины; $u = \{u_1, u_2, \ldots, u_N\}$ — искомая вектор-функция; $f(x,u) = \{f_1(x,\vec{u}), \ldots, f_N(x,\vec{u})\}$ — вектор правых частей. Относительно задачи (1-2) будем предполагать выполненными достаточные условия существования на отрезке $|x - x_0| < a$ решения $u(x)$ задачи (1)-(2).

Эйлеру принадлежит идея и рассмотрение простейшего численного метода, основанного на возможности получить разложение по формуле Тейлора для искомого решения $u(x)$ в окрестности точки x_n

$$u_{n+1} = u(x_{n+1}) = u_n + h_n u'_n + \frac{1}{2}h_n^2 u''_n + \cdots + \frac{h_n^s}{(s)!}u_n^{(s)} + O(h_n^{(s+1)}), \tag{3}$$

где $h_n = x_{n+1} - x_n$. При этом необходимые производные функции $u(x)$ можно найти дифференцируя в силу уравнения (1) функцию $f(x, u(x))$ нужное число раз

$$u' = f(x,u); \quad u'' = \frac{d}{d\,x} f(x,u(x)) = f_x + f_u \cdot \underbrace{u_x}_{\equiv f(x,u)} = f_x + f f_u, \quad \text{и т.д.} \tag{4}$$

Однако использовать разложение (3) с большим числом членов невыгодно: и из-за громоздкости формул (4), и из-за того, что, как правило, правая часть в (1) известна лишь приближённо и её явное численное дифференцирование нежелательно.

1.2 Метод Рунге-Кутта

Идея Рунге метода Рунге-Кутта состоит в том, чтобы используя метод неопределённых коэффициентов аппроксимировать с тем же порядком точности $O(h_n^s)$ многочлен Тейлора в формуле (3). Представим приращение функции $u(x)$ в точке x_n в виде

$$\Delta u(x_n) = u(x_{n+1}) - u(x_n) = h_n \left(\underbrace{u'_n + \frac{1}{2} h_n u''_n + \cdots + \frac{h_n^{s-1}}{(s)!} u_n^{(s)}}_{P_{s-1}(h_n)} + O(h_n^{(s)}) \right).$$

Обозначим текущий шаг $h_n \equiv h$. Речь идёт об аппроксимации многочлена

$$P_{s-1}(h) = u'_n + \frac{1}{2} h u''_n + \cdots + \frac{h^{s-1}}{(s)!} u_n^{(s)}$$

с порядком $O(h^s)$. Ограничимся рассмотрением простейшего случая $s = 2$. Тогда у многочлена первого порядка

$$P_1(h) = u'_n + \frac{h}{2} u''_n = f(x_n, u_n) + \frac{h}{2} \frac{d}{d\,x} f(x,u)|_{(x_n,u_n)}$$

необходимо со вторым порядком аппроксимировать производную u''_n.

Пусть $y(x)$ — приближенная функция, дающая такую аппроксимацию. Для аппроксимации производной df/dx мы используем разностное отношение $[f(\tilde{x}, \tilde{y}) - f(x,y)]/\Delta x$ с неопределенными пока $\tilde{x}, \tilde{y}$. В таком случае приращение функции y имеет вид

$$\Delta y_n = y_{n+1} - y_n = h \left\{ \beta f(x_n, y_n) + \alpha f(x_n + \gamma h, y_n + \delta h) \right\}.$$

Здесь α, β, γ и δ — параметры, значения которых нужно определить.

Разложим полученное приращение Δy_n в ряд по степеням h, получим

$$y_{n+1} = y_n + h(\alpha + \beta) f(x_n, y_n) + \alpha h^2 (\gamma f_x + \delta f_u)|_{(x_n,y_n)} + O(h^3). \tag{*}$$

Выберем параметры α, β, γ и δ так, чтобы разложение для функции y с тем же порядком аппроксимировало разложение истинного решения u. Для этого приравнивая

коэффициенты в главных порядках по h полученной формулы (*) и формулы (3), найдём

$$\alpha + \beta = 1, \quad \alpha\gamma = \frac{1}{2}, \quad \alpha\beta = \frac{1}{2} f(x_n, y_n).$$

Выражая все параметры через α, получим однопараметрическое семейство двучленных схем Рунге-Кутта второго порядка точности

$$y_{n+1} = y_n + h\left[(1-\alpha) f(x_n, y_n) + \alpha f\left(x_n + \frac{h}{2\alpha}, y_n + \frac{h}{2\alpha} f_n\right)\right], \tag{5}$$

где $0 < \alpha \leq 1$.

Замечания:

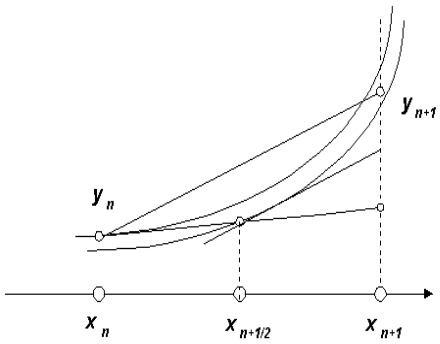
1) Выбрать параметр α так, чтобы схема (5) давала бы аппроксимацию третьего порядка невозможно.

2) Приведем без доказательства теорему. *Если $f(x, u)$ непрерывна и ограничена вместе со своими вторыми производными, то решение, полученное по схеме (5), равномерно сходится к точному решению с погрешностью $O(\max h_n^2)$, т.е. двучленная схема Рунге-Кутта имеет второй порядок точности.*

3) Формула (5) используется на практике обычно либо при $\alpha = 1$, либо при $\alpha = 1/2$. При $\alpha = 1$ схема имеет особенно простой вид

$$y_{n+1} = y_n + \frac{h}{2} f\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}h f_n\right). \tag{6}$$

Поясним её смысл. Сначала, вычислив наклон интегральной кривой уравнения (1) $f_n = f(x_n, y_n)$, делаем половинный шаг по схеме ломанных, т.е. по касательной данного наклона, и находим



$$y_{n+1/2} = y_n + \frac{1}{2} h f_n.$$

Затем в найденной точке определяем наклон интегральной кривой $y'_{n+1/2} = f(x_{n+1/2}, y_{n+1/2})$. По этому наклону определяем приращение функции на целом шаге

$$y_{n+1} = y_n + h y'_{n+1/2}.$$

Схемы подобного типа называют "предиктор-корректор".

Задача. *Дать аналогичную интерпретацию слуяаю схемы с $\alpha = 1/2$.*

Метод Рунге-Кутта позволяет строить схемы различного порядка точности. При аппроксимации многочлена Тейлора второго порядка

$$P_2(h) = u'_n + \frac{h}{2} u''_n + \frac{h^2}{3!} u'''_n$$

с точностью $O(h^3)$ получают наиболее употребительную схему четвёртого порядка точности (точнее семейство четырёхчленных схем указанного порядка точности)

$$y_{n+1} = y_n + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4),$$

$$k_1 = f(x_n, y_n), \quad k_2 = f(x_n + \frac{h}{2}, y_n + \frac{h}{2}k_1), \tag{7}$$

$$k_3 = f(x_n + \frac{h}{2}, y_n + \frac{h}{2}k_2), \quad k_4 = f(x_n + h, y_n + hk_3).$$

Схемы Рунге-Кутта обладают важными достоинствами: все они имеют хорошую точность; они являются явными; допускают расчет с переменным шагом; легко обобщаются на случай систем дифференциальных уравнений. Имеено эти свойства особенно ценны при расчетах на ЭВМ.

Рекомендации:

1) Если правая часть дифференциального уравнения (1) ограничена вместе со своими производными до четвёртого порядка, то схема (7) дает хорошие результаты благодаря малому коэффициенту в остаточном члене и быстрому возрастанию точности схемы при уменьшении шага. Если же указанных производных у правой части нет. то не худшую точность имеют схемы и меньшего порядка точности (5).

2) Шаг сетки при расчетах следует выбирать настолько малым, чтобы обеспечить требуемую точность расчета. Других, ограничительных условий на шаг схемы в методе Рунге-Кутта нет.

3) Выражения остаточных членов для формул Рунге-Кутта достаточно громоздки, поэтому трудно получить априорную оценку точности метода, однако, проводя расчеты на сгущающихся сетках, можно дать апостериорную оценку точности по методу Рунге.

ГЛАВА VIII

ЭЛЕМЕНТЫ ТЕОРИИ РАЗНОСТНЫХ СХЕМ

§1. Метод конечных разностей в прикладных задачах

1.1 Общая постановка задачи

Универсальным методом приближённого решения, применимым для широкого круга задач математической физики, является метод конечных разностей. Как правило задачи матаматической физики представляют собой системы нелинейных уравнений в частных производных, рассматриваемых в некоторой t-цилиндрической области D:

$$\bar{D} = \{(x, y, z; t) : (x, y, z) \in \bar{G}, t \in [t_\circ, T]\} = \bar{G} \times [t_\circ, T].$$

При этом естественным образом выделяется "эволюционный" характер переменной t. Решение интересующей нас задачи подчинено в $\bar{D}$ дополнительным требованиям: 1) условия при $t = t_0$ (на гиперплоскости $t = t_0$) называются *начальными условиями*; 2) условия на границе $\partial D \equiv \gamma$ области D — *краевыми* или *граничными* условиями.

Задача с *начальными условиями* – задача в неограниченной области D – называется задачей Коши; в отличии от *краевой* или *смешанной краевой* задачи.

Удобна общая постановка задачи, не связанная с выделением одной из переменных. Пусть $(x_1, ..., x_p) \equiv x \in D : \partial D = \Gamma$. Тогда для интересующей нас функции $u(x)$ имеем задачу:

$$A[u(x)] = f(x), \quad x \in D \tag{1}$$
$$R[u(x)] = \mu(x), \quad x \in \Gamma, \tag{2}$$

где A и R –дифференциальные операторы задачи и краевых условий. Относительно задачи (1-2) будем предполагать ,что она поставлена корректно, то есть операторы A и R; область D и её границы Γ таковы, что при выборе соответствующих классов функций и правых частей в уравнениях (1) и (2) решение существует, единственно и непрерывно зависит от начальных данных [*1)].

С точки зрения приложений нас, естественно, будет интересовать случай, когда оператор A — линейный дифференциальный оператор в частных производных второго порядка (согласно обычной классификации уравнений это — эллиптическое, гиперболическое или параболическое уравнение). Хотя, конечно, задача (1-2) может быть и другой природы.

[*1)] И коэффициентов уравнения, то есть соответствующих операторов задачи (1-2).

1.2 Разностная схема

Введём в области $\bar{D} = D + \Gamma$ сетку $\Omega_h = x_{i\in I}$ состоящую из множества внутренних узлов ω_h и множества граничных узлов Γ_h:

$$\Omega_h = \{x_i\}_I = \omega_h \cup \Gamma_h.$$

Мы пока абстрагируемся от способа конкретного получения сетки Ω_h в области $\bar{D}$; смысла параметра "h" в соответствующих сетках, контролирующего как пространственные, так и временные размеры сетки; особенностей получения сетки Γ_h на границе области — оставим эти вопросы до рассмотрения конкретных задач.

Далее, рассмотрим сеточные функции $y(x) \equiv y_h(x),\ x \in \Omega_h$ дискретного переменного $\{x_i\}$ и с их помощью построим приближенное решение задачи (1-2). Для этого относительно $y_h(x)$ сформулируем "разностную задачу", обычно "заменяя" операторы исходной задачи A и R и их сеточными аналогами A_h и R_h. Тогда на сеточном шаблоне $\Omega_h = \omega_h \cup \Gamma_h$ имеем

$$A_h y_h(x) = \varphi_h(x), \quad x \in \omega_h \tag{3}$$
$$R_h y_h(x) = \chi_h(x), \quad x \in \Gamma_h, \tag{4}$$

Задачу (3)-(4) назовём *разностной схемой* для задачи (1)-(2). Обычно это алгебраическая система уравнений относительно $y_i = y_h(x_i)$.

При переходе от исходной задачи (1)-(2) к её разностному аналогу (3)-(4) особенно важны три группы вопросов:

–существование, единственность и алгоритм построения разностного решения y_h ;

–при каких условиях разностное решение $y_h(x)$ стремится к точному решению $u(x)$ и какова при этом скорость сходимости;

–из каких соображений и как конкретно выбирать сетку Ω_h и строить разностную схему: A_h, R_h и φ_h, χ_h в задаче (3)-(4).

§2. Основные понятия и теоремы теории разностных схем

2.1 Невязка разностной схемы.

При построении разностного уравнения задачи

$$A[u] = f \quad \Rightarrow \quad A_h y = \varphi_h$$

мы получили задачу, которой точное решение $u(x)$, как правило, не удовлетворяет (мы подразумеваем простейшую схему проектирования $u(x)$ на сетку Ω_h $\{u(x_i)\}$).

Сеточную функцию

$$\psi_h = \varphi_h - A_h u$$

называют *невязкой* сеточного уравнения (3). Её удобно представить на решении $u(x)$ в виде

$$\psi_h = (Au - f)_h - (A_h u - \varphi_h) \quad \text{на } \omega_h. \tag{5}$$

Аналогично определяются невязки граничных условий

$$\eta_h(x) = (Ru - \mu)_h - (R_h u - \chi_h) \quad \text{на } \Gamma_h. \tag{5'}$$

Как правило невязки $\psi_h(x)$ и $\eta_h(x)$ оценивают по параметру h через разложение в ряд Тейлора в предположении достаточной гладкости соответствующего решения $u(x)$ для получения представления невязки с остаточным членом вида $O(h^n)$.

2.2 Аппроксимация разностной схемы

Разностная схема (3)-(4) аппроксимирует задачу (1)-(2), если имеет место:

$$\| \psi_h(x) \|_{\varphi_h} \to 0, \quad \| \eta_h(x) \|_{\chi_h} \to 0 \quad \text{при } h \to 0 \tag{6}$$

То есть соответствующие невязки стремяться к нулю при $h \to 0$.
Аппроксимация задачи (1)-(2) имеет порядок k ,если

$$\| \psi_h(x) \|_{\varphi_h} = O(h^k); \quad \| \eta_h(x) \|_{\chi_h} = O(h^k), \quad h \to 0. \tag{6'}$$

В этих определениях нормы вычисляются для сеточных функций на ω_h и Γ_h, но в своих функциональных пространствах (соответствующих правых частей). Вопрос о выборе норм отложим до рассмотрения частных задач. Обычно это сеточные аналоги чебышевской нормы в C или гильбертовой нормы в L_2.
Замечания:

Само решение задачи (1)-(2) ,как правило, неизвестно и использовать его для получения невязок ψ_h и η_h затруднительно. Поэтому берут достаточно широкий класс функций $\mathcal{V}$ и требуют аппроксимации порядка k задачи (1)-(2) $\forall v \in \mathcal{V}$, то есть

$$\| (Av - f)_h - (A_h v - \varphi_h) \|_{\varphi_h} = O(h^k), \quad h \to 0.$$

При этом на решении $v \equiv u(x)$ задачи (1)-(2) аппроксимация будет не хуже ,чем порядка k (а может быть и лучше).

Как правило схема (3)-(4) по различным переменным имеет различные порядки аппроксимации, например, невязка уравнения

$$\| \psi_h \|_{\varphi_h} = O(\tau^p + h^k), \quad \text{при } \tau \to 0, \quad h \to 0.$$

Такая аппроксимация называется *абсолютной* в отличии от *условной* аппроксимации в случае, когда, например

$$\| \psi_h \|_{\varphi_h} = O(\tau^p + h^k + \frac{\tau^r}{h^\delta}), \quad \tau \to 0, \quad h \to 0, \quad \frac{\tau^r}{h^\delta} \to 0.$$

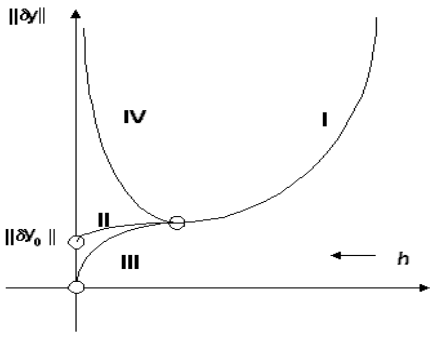
При *условной аппроксимации* разностное уравнение может аппроксимировать различные дифференциальные задачи.

2.3 Устойчивость разностной схемы

Отсутсвие устойчивости разностной схемы характеризуется тем, что малые ошибки, допущенные на каком-либо этапе вычисления, в дальнейшем сильно возрастают и делают непригодным результат расчёта (или вообще невозможным сам расчёт). Обычно устойчивость разностной схемы оценивают по погрешности входных данных, поскольку погрешность аппроксимации, в силу определения (6), при $h \to 0$ стремится к нулю. Выделим в структуре погрешности эти слагаемые:

$$\delta y_h = \delta y_h^{in.} + \delta y_h^{app.}.$$

Типичный график зависимости погрешности сеточного решения от величины шага таков:



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 расчёта.

Если же схема неустойчива (IV), то при $h \to 0$ погрешность $\|\delta y_h\|$ возрастает (ибо растёт объём неустойчивых вычислений). Погрешность $\|\delta y_h\|$ будет иметь ненулевой минимум и уже невозможно получить сколь угодно высокую точность расчёта.

Как правило погрешности *входных данных* и *аппроксимации* имеют степенной характер зависимости от $h \Rightarrow h^\alpha$; а неустойчивость приводит к возрастанию погрешности решения по экспоненциальному закону $\sim b^{a/h}$ и при $h \to 0$ расчёт теряет смысл. Напомним

Разностная схема (3-4) устойчива по входным данным φ и χ, если решение разностной схемы непрерывно зависит от входных данных и эта зависимость равномерна относительно шага сетки h, то есть $\forall \varepsilon > 0\ \exists \delta(\varepsilon) > 0$ (δ не зависит от h) такое, что

$$\begin{array}{l} \forall \chi_1, \chi_2: \quad \| \chi_1 - \chi_2 \| < \delta(\varepsilon) \\ \forall \varphi_1, \varphi_2: \quad \| \varphi_1 - \varphi_2 \| < \delta(\varepsilon) \end{array} \quad \Rightarrow \quad \| y_1(x) - y_2(x) \|_{y_h} < \varepsilon. \tag{7}$$

Для линейных схем разностное решение линейно зависит от входных данных (в силу линейности обратного оператора)и $\delta(\varepsilon) = C\varepsilon$. Тогда

$$\|y_1 - y_2\| \le C_1 \|\varphi_1 - \varphi_2\|_{\varphi_h} + C_2 \|\chi_1 - \chi_2\|_{\chi_h}. \tag{7'}$$

Замечания:

На устойчивость разностной схемы влияет не только аппроксимация уравнений (1) (то есть оператора A), но, и особенно, краевых условий (2).

Если переменных в задаче несколько, то рассматривают безусловную и условную устойчивость;

Входное значение $\chi_h(x)$ на гиперплоскости $t = t_\circ$ выделяют особо,и соответствующая устойчивость называется устойчивостью по начальным условиям. Тут важна особая роль t. Мы ограничимся рассмотрением разностных схем, в которых сеточная функция рассматривается на двух временных слоях t_m; t_{m+1}, то есть $y \equiv y_h(x;t_m)$ и $\hat{y} \equiv y_h(x;t_{m+1})$. Общий вид такой схемы:

$$B_h \frac{\hat{y}-y}{\tau} + A_h y = \varphi_h.$$

Для такой схемы решение смешанной задачи Коши (с краевыми условиями) на некотором слое t^* можно рассматривать как начальное условие для всех последующих слоёв по t.

Определение: *Двуслойная схема называется равномерно устойчивой по начальным данным, если при постановке начальных данных на любом слое t^*, $(t_0 \le t^* < t < T)$ она по ним устойчива, причём эта устойчивость равномерна по t^*.*

Для линейных разностных схем это означает, что $\exists C > 0$ не зависящее от t^* и h и

$$\| y_1(t) - y_2(t) \|_{y_h} \le C \| y_1(t^*) - y_2(t^*) \|, \quad t_0 \le t^* < t < T \qquad (7'')$$

где $y_1(x;t), y_2(x;t)$ — решение разностной задачи с одинаковой правой частью $A_h y = \varphi_h$, но различными начальными данными $\chi_{1,2}|_{t^*}$.

Из равномерной устойчивости $(7'')$ следует $(7')$ (но не наоборот).

Теорема 1. (достаточный признак равномерной устойчивости):

Пусть $y_1(x;t)$ и $y_2(x;t)$ решения разностной задачи $A_h y = \varphi_h$ с одинаковой правой частью, отвечающие различным начальным условиям $\chi_{1,2}|_{t^*=t_\circ}$. Тогда для равномерной устойчивости $\{A_h; R_h\}$ по начальным данным достаточно, чтобы для всех слоёв по t имело место

$$\| \hat{y}_1 - \hat{y}_2 \|_{y_h} \le (1 + C\tau) \| y_1 - y_2 \|_{y_h}, \quad C \ge 0 \qquad (8)$$

Доказательство: Если на некотором слое t_* в решении содержится ошибка δy, то при переходе на следующий слой она возрастает не больше чем в $(1 + C\tau) \le e^{C\tau}$ раз. При достижении слоя T за $\frac{T-t^*}{\tau}$ шагов ошибка вырастет не более чем в $e^{C(T-t^*)}$ раз, то есть не более чем в $e^{C(T-t_0)}$ раз. Следовательно

$$\| \delta y \| \le A \| \delta y(t_0) \| .$$

Эта оценка равномерна по t^* и h ■

Фактический рост погрешности не более чем в $(1 + C\tau)^{\frac{T-t_0}{\tau}}$ раз.

Теорема 2. (признак устойчивости двуслойной разностной схемы по правой части):

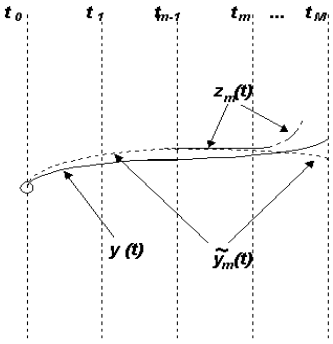
Пусть двуслойная разностная схема $A_h y = \varphi_h$ равномерно устойчива по начальным данным и такова, что если два её решения $A_h y_{1,2} = \varphi_{1,2}$ на некотором слое t_m равны $y_1(x;t_m) = y_2(x;t_m)$, то на следующем слое t_{m+1} выплнено соотношение

$$\| \hat{y}_1 - \hat{y}_2 \| \le C\tau \| \varphi_1 - \varphi_2 \|, \quad C > 0,$$

$C-const$ (не зависит от h), в таком случае разностная схема устойчива по правой части φ_h.

Доказательство: Итак, пусть возмущение связанно только с правой частью φ. Тогда пусть $y(x;t)$ — решение невозмущённой разностной задачи $A_h y = \varphi$; $\tilde{y}(x;t)$ — решение возмущённой разностной задачи $A_h\tilde{y} = \tilde{\varphi}$, причём $y(t_0) = \tilde{y}(t_0)$ (ибо нас интересует только возмущение правой части).

Введём в рассмотрение последовательность сеточных функций $\{z_m(x;t)\}_{m=1,2,...}$, определенных при $t \geq t_{m-1}$ из условий:



$$\begin{cases} z_m(t_{m-1}) = z_{m-1}(t_{m-1}) \\ A_h z_m = \begin{cases} \tilde{\varphi} & , \quad t_{m-1} < t \leq t_m \\ \varphi & , \quad t > t_m \end{cases} \\ z_1(t_0) = y(t_0) = \tilde{y}(t_0) = z_0(t_0) \end{cases}$$

На каждом из слоёв $t \in [t_{m-1}, t_m]$ решение возмущённой 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 попадём в начальную по t точку. В точке $t = t_{m-1}$ и $\tilde{y}$ и z_{m-1} имеют то же начальное условие и на интервале $(t_{m-1}; t_m)$ удовлетворяют возмущенной задаче $A_h(\cdot) = \tilde{\varphi}$.

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:

1) $\| z_{m+1}(t_{m+1}) - z_m(t_{m+1}) \| \leq C\tau \| \varphi - \tilde{\varphi} \|_\varphi$.

2) В силу равномерной устойчивости нашей задачи по начальным данным при $t \geq t_{m+1}$ функции $z_{m+1}(t)$ и $z_m(t)$ удовлетворяют одному уравнению но разностным начальным условиям. В таком случае на последнем временном слое t_M получим:

$$\| z_{m+1}(t_M) - z_m(t_M) \| \leq C_2 \| z_{m+1}(t_{m+1}) - z_m(t_{m+1}) \| \leq C_2 C\tau \| \varphi - \tilde{\varphi} \| .$$

Откуда:

$$\| z_M(t_M) - z_0(t_M) \| \leq \| z_M - z_{M-1} \| + \| z_{M-1} - z_{M-2} \| + ... + \| z_1 - z_0 \| \leq$$

$$\leq MC_2 C\tau \| \varphi - \tilde{\varphi} \| = A(T - t_\circ) \| \varphi - \tilde{\varphi} \| .$$

Таким образом имеет место устойчивость разностной схемы по правым частям ■

Замечание: Сформулируем без доказательства достаточные условия устойчивости двуслойной разностной схемы

$$B\frac{\hat{y} - y}{\tau} + Ay = \varphi.$$

Если A и $B > 0$, причм $B \geq \frac{\tau A}{2} > 0$, то $\| \hat{y} \|_A \leq \| y \|_A$, то есть схема устойчива в A-энергетической норме по начальным данным.

2.4 Сходимость разностной схемы

Решая сеточную задачу (3)-(4) нас естественно интересует близость сеточного решения $y(x)$ к решению $u(x)$ задачи (1)-(2). *Разностное решение $y(x)$ сходится к решению $u(x)$, если*

$$\| y(x) - u(x) \|_h \to 0 \quad \text{при } h \to 0. \tag{10}$$

Разностное решение имеет порядок точности k, если

$$\| y(x) - u(x) \| = O(h^k), \quad h \to 0. \tag{10'}$$

(или обладает сходимостью порядка k).

Напомним ещё раз, что мы рассматриваем лишь корректные разностные схемы (3)-(4), то есть решение разностной схемы существует и единственно при любых входных данных φ и χ из заданных классов функций и схема устойчива по входным данным (её решение непрерывно от них зависит).

Теорема3: *Если решение задачи (1)-(2) $u[f, \mu]$ существует, разностная схема (3)-(4) корректна и аппроксимирует задачу (1)-(2), то разностное решение $y[\varphi, \chi]$ сходится к точному:*

$$\lim_{h \to 0} \| y_h - u \| = 0.$$

("Аппроксимация + Устойчивость $\Rightarrow$ Сходимость").

Доказательство: Запишем невязку разностной схемы (3)-(4).

$$\begin{array}{l} \psi_h = (Au - f)_h - (A_h u - \varphi_h) = \varphi_h - A_h u \\ \eta_h = (Ru - \mu) - (R_h u - \chi_h) = \chi_h - R_h u \end{array} \quad \Leftrightarrow \quad \begin{array}{l} A_h u = \varphi_h - \psi_h \\ R_h u = \chi_h - \eta_h. \end{array} \tag{$*$}$$

Функция $u(x)$ удовлетворяет задаче ($*$) — возмущённой задаче (3)-(4). Так как схема устойчива, то $\forall \varepsilon > 0, \exists \delta(\varepsilon) > 0$

$$\| \psi_h \|_{\varphi_h} < \delta(\varepsilon), \quad \| \eta_h \|_{\chi h} < \delta(\varepsilon) \quad \Rightarrow \quad \| y - u \|_{y_h} < \varepsilon.$$

В силу аппроксимации $\forall \delta > 0$, $\exists h_0$, $\forall h < h_0$ имеет место

$$\| \psi_h \|_{\varphi} < \delta, \quad \| \eta \|_{\chi} < \delta.$$

Таким образом: $\forall \varepsilon > 0$, $\exists h_0(\delta(\varepsilon))$, $\forall h < h_0$ имеем

$$\| y_h - u \|_{y_h} < \varepsilon,$$

то есть $y \to u$ при $h \to 0$ ■

Замечания:

Если какое-либо данное нам условие аппроксимировано точно, то устойчивость по ним можно не требовать, так как они не вносят погрешности в решение (кроме ошибок округления, тогда устойчивость по этим данным нужна).

Для условной аппроксимации (или устойчивости) сходимость тоже носит условный характер.

Для линейных разностных схем имеет место:

Теорема 4. *Пусть выполнены условия Теоремы 1, схема A_h, R_n линейна и имеет порядок аппроксимации k, то схема (3)-(4) сходится и её точность (сходимость)*

не ниже порядка k (порядка аппроксимации).
Доказательство: Рассмотрим погрешность разностного решения

$$z(x) = y(x) - u(x).$$

Мы получили для решения исходной задачи разностную схему, возмущённую невязками

$$\begin{cases} A_h u = \varphi - \psi, & x \in \omega_h \\ R_h u = \chi - \eta, & x \in \Gamma_h \end{cases}$$

Вычитая эти уравнения из соответствующих уравнений (3)-(4), найдём:

$$\begin{cases} A_h z = \psi \\ R_h z = \eta \end{cases} \qquad (**)$$

Схема (**) устойчива, то есть

$$\| z \|_{y_h} \leq C_1 \| \psi \|_{\varphi} + C_2 \| \eta \|_{\chi} .$$

Но, поскольку исходная схема (3)-(4) обладает аппроксимацией порядка k, то

$$\| z \|_{y_h} \leq C_1 \alpha h^k + C_2 \beta h^k = C h^k.$$

Фактическая сходимость может иметь более высокий порядок.

§3. Разностные схемы для одномерного уравнения теплопроводности

3.1 Постановка задачи. Разностная схема

Рассмотрим задачу о распространении тепла на отрезке в случае простейших краевых условий 1-го рода (условий Дирихле)

$$u_t = a^2 u_{xx} + f(x,t), \qquad 0 < x < l, \quad t > 0$$

начальные условия

$$u(x,0) = \mu_1(x) \equiv \mu(x) \qquad (11)$$

однородные краевые условия

$$u(0,t) = \mu_2(t) \equiv 0; \quad u(l,t) = \mu_3(t) \equiv 0, \quad t \geq 0.$$

а) Конечно-разностная аппроксимация простейших дифференциальных операторов первого порядка.

Введем в области $\bar{D} = [0; l] \times [0; T]$ сетку $\Omega = \omega_h \times \omega_\tau$, где

$$\omega_h = \left\{ \begin{array}{c} 0 = x_0 < x_1 < \cdots < x_N = l \\ x_n = x_0 + nh, \quad h = \frac{x_N - x_0}{N} \end{array} \right\} \quad \text{и} \quad \omega_\tau = \left\{ \begin{array}{c} 0 = t_0 < t_1 < \cdots < t_M = T \\ t_m = t_0 + m\tau, \quad \tau = \frac{T - t_0}{M} \end{array} \right\}.$$

Рассмотрим сеточную функцию $y(x_n; t_m) = y_n^m = y$ на сетке $\Omega \equiv \omega_{h,\tau}$. Построим сеточные аналоги простейших дифференциальных операторов первого порядка:

$$\hat{L}u = u'(x) = \frac{du}{dx} \quad \Rightarrow \quad \begin{array}{ll} l_x y = y_{x,i} = \frac{y_{i+1} - y_i}{h}, & \text{производная вперёд} \\ l_{\bar{x}} y = y_{\bar{x},i} = \frac{y_i - y_{i-1}}{h}, & \text{производная назад} \\ l_{\overset{0}{x}} y = y_{\overset{0}{x},i} = \frac{y_{i+1} - y_{i-1}}{2h}, & \text{центральная производная} \end{array} \tag{12}$$

Их аппроксимация $L_h u - (Lu)_h$ имеет следующий порядок:
Для производной вперёд l_x

$$\frac{u_{i+1} - u_i}{h} - u'(x_i) = \frac{u(x_i + h) - u_i}{h} - u'(x_i) =$$

$$= \frac{u(x_i) + u'(x_i)h + O(h^2) - u_i}{h} - u'(x_i) = O(h),$$

т.е. обладает аппроксимацией 1-го порядка.
Аналогично $l_{\bar{x}}$

$$\frac{u_i - u_{i-1}}{h} - u'(x_i) = \frac{u(x_i) - u(x_i - h)}{h} - u'(x_i) =$$

$$= \frac{u_i - [u(x_i) - u'(x_i)h + O(h^2)]}{h} - u'(x_i) = O(h).$$

Центральная производная $l_{\overset{0}{x}}$ имеет повышенный порядок аппроксимации

$$\frac{u_{i+1} - u_{i-1}}{2h} - u'(x_i) = \frac{u(x_i + h) - u(x_i - h)}{2h} - u'(x_i) =$$

$$= \frac{u(x_i) + u'(x_i)h + h^2/2\ u''(x_i) + O(h^3) - [u(x_i) - u'(x_i)h + h^2/2\ u''(x_i) + O(h^3)]}{2h} -$$

$$- u'(x_i) = O(h^2).$$

b) Конечно-разностная аппроксимация простейших дифференциальных операторов второго порядка.

Определим вторую разностную производную для узла x_i (рекурентно):

$$\hat{L} = \frac{d^2}{dx^2} = \frac{d}{dx}\left(\frac{d}{dx}\right) \quad \Rightarrow \quad y_{\overline{x}x,i} = (y_{\overline{x}})_{x,i} = \frac{1}{h}\left(y_{\overline{x},i+1} - y_{\overline{x},i}\right) =$$

$$= \frac{1}{h}\left(\frac{y_{i+1} - y_i}{h} - \frac{y_i - y_{i-1}}{h}\right) = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} = y_{x\overline{x},i}. \tag{13}$$

Получим её порядок аппроксимации

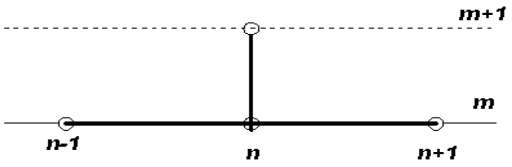
$$L_h u - (Lu)_h = u_{\overline{x}x,i} - \left(u''\right)_i = \frac{u(x_i + h) - 2u(x_i) + u(x_i - h)}{h^2} - u''(x_i) =$$

$$= \frac{1}{h^2}\left(h^2 u''(x_i) + O\left(h^4\right)\right) - u''(x_i) = O\left(h^2\right).$$

Аналогично мы можем построить аппроксимации и более сложных производных.

Разностная схема. После аппроксимации простейших дифференциальных операторов, вернемся к уравнению (11.1).

Используя так называемый метод *разностной аппроксимации*, мы можем каждый из дифференциальных операторов задачи (11) аппроксимировать соответствующим разностным оператором(12), (13). Производная вперед по t для (n, m)-го узла



$$y_{t;n,m} = \frac{y_n^{m+1} - y_n^m}{\tau} = \frac{\hat{y}_n - y_n}{\tau} = \frac{\hat{y} - y}{\tau}.$$

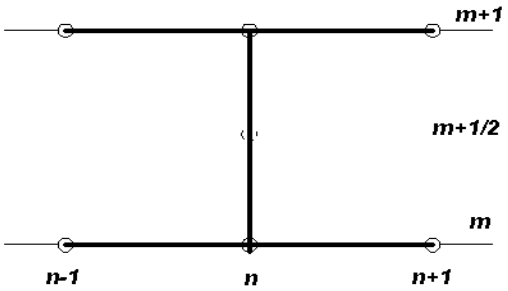
Это выражение рассматривается относительно текущего узла x_n на двух слоях по t.

Пространственные производные второго порядка аппроксимируются разностным оператором

$$y_{\overline{x}x;n,m} = \frac{1}{h^2}\left(y_{n+1}^m - 2y_n^m + y_{n-1}^m\right) = \Lambda_h y_n^m \equiv \Lambda_h y.$$

При построении такой разностной аппроксимации на $\omega_{h,\tau}$ мы использовали шаблон из четырех узлов.

Относительно $(m+1)$-го временного слоя схема получилась *явной* — с $(m+1)$-го временного слоя используется только одно значение сеточной функции. В дальнейшем мы покажем, что простейшая явная схема не является наилучшей в смысле аппроксимации и, особенно, устойчивости. Поэтому сразу же рассмотрим однопараметрическое семейство схем на шеститочечном шаблоне:



$$\frac{y_n^{m+1}-y_n^m}{\tau}=\alpha^2\Lambda_{\overline{x}x}\left\{\sigma y_n^{m+1}+(1-\sigma)\,y_n^m\right\}+\varphi_n^m,$$

где вес $\sigma\in[0;1]$.

При $\sigma=0$ получается чисто *явная* схема, при $\sigma=1$ — чисто *неявная* схема. При аппроксимации правой части $f(x,t)\Rightarrow\varphi_n^m$ мы использовали, так называемый, *метод неопреленных коэффициентов* в простейшей его форме, когда подбирается всего один коэффициент φ_n^m (без дополнительного сложного шаблона).

Итак, получаем разностную задачу:

$$\begin{cases}\frac{1}{\tau}\left(y_n^{m+1}-y_n^m\right)=\alpha^2\Lambda_h\left\{\sigma\hat{y}+(1-\sigma)\,y\right\}+\varphi_n^m; & (x_n,t_m)\in\omega_{h,\tau}\\ y_n^0=\chi_n;\\ y_0^m=y_N^m=0=\chi_{0,N}^m\end{cases}\tag{14}$$

Уравнениие (14.1) записано относительно внутренних узлов (n,m) сетки $\overline{\Omega}$. При аппроксимации начальных и краевых условий мы также использовали метод неопределенных коэффициентов. Теперь изучим свойства построенной разностной схемы.

3.2 Порядок аппроксимации разностной схемы (14)

Напомним еще раз, что для определения порядка аппроксимации разностной схемы (14), нужно точное решение (11) подставить в эту схему и, в предположении достаточной гладкости решения $u(x,t)$, определить порядки невязок ψ и η по h и τ.

Одновременно с этим, мы проследим идею метода неопределенных коэффициентов, выбираемых из соображений обеспечения максимального порядка аппроксимации (на примере построения φ_n^m и частично χ_n).
Введем в рассмотрение промежуточный слой по t :

$$\overline{t}=t_m+\frac{\tau}{2}=t_{m+\frac{1}{2}}=m\tau+\frac{\tau}{2}.$$

Тогда

а) временна́я часть:

$$\frac{1}{2\frac{\tau}{2}}\left(u_n^{m+1}-u_n^m\right)=u_{0\atop t;n,m+\frac{1}{2}}=u_t'\left(x_n,t_{m+\frac{1}{2}}\right)+O\left(\tau^2\right);$$

б) пространственная часть:

$$\sigma\hat{u}+(1-\sigma)\,u=\sigma u\left(x_n,t_{m+\frac{1}{2}}+\frac{\tau}{2}\right)+(1-\sigma)\,u\left(x_n,t_{m+\frac{1}{2}}-\frac{\tau}{2}\right)=$$

$$=\sigma\left\{u\left(x_n,t_{m+\frac{1}{2}}\right)+\frac{\tau}{2}\overline{u}_t+\frac{1}{2!}\left(\frac{\tau}{2}\right)^2\overline{u}_{tt}+O\left(\tau^3\right)\right\}+$$

$$+(1-\sigma)\left\{u\left(x_n,t_{m+\frac{1}{2}}\right)-\frac{\tau}{2}\overline{u}_t+\frac{1}{2!}\left(\frac{\tau}{2}\right)^2\overline{u}_{tt}+O\left(\tau^3\right)\right\}=$$

$$=\overline{u}+\tau\left(\sigma-\frac{1}{2}\right)\overline{u_t}+O\left(\tau^2\right).$$

Здесь чертой сверху обозначено значение функции в точке $(x_n;t_{m+1/2})$.
Следовательно,

$$\Lambda_h\left[\sigma\,\hat{u}+(1-\sigma)\,u\right]=\Lambda_h\left[\overline{u}+\tau\left(\sigma-\frac{1}{2}\right)\overline{u}_t+O\left(\tau^2\right)\right]=$$

$$=\overline{u}_{xx}+\left(\sigma-\frac{1}{2}\right)\tau\overline{u}_{txx}+O\left(\tau^2+h^2\right).$$

Таким образом подстановка $u(x,t)$ в разностное уравнениие (14.1) дает

$$\underline{u_t\left(x_n,\overline{t}\right)}+O\left(\tau^2\right)=\underline{a^2u_{xx}\left(x_n,\overline{t}\right)}+a^2\tau\left(\sigma-\frac{1}{2}\right)u_{txx}\left(x_n,\overline{t}\right)+\varphi_n^m+O\left(\tau^2+h^2\right).$$

В силу задачи (11) подчеркнутые члены анулируются, если в уравнении есть слагаемое $f\left(x_n,\overline{t}\right)$. Таким образом, если мы хотим обеспечить аппроксимацию задачи (11), необходимо:

$$\varphi_n^m=f\left(x_n,\overline{t}\right)=f\left(x_n,t_{m+\frac{1}{2}}\right).$$

Тогда:

1) при $\sigma\neq\frac{1}{2}$ мы получаем аппроксимацию уравнения (11.1) с порядком $O\left(\tau+h^2\right)$;

2) при $\sigma=\frac{1}{2}$ мы получаем повышенный порядок аппроксимации $O\left(\tau^2+h^2\right)$ (обратим внимание на наличие симметрии в сеточном шаблоне).

3) Аппроксимация начальных условий в этой задаче тривиальна:

$$\chi_n^0=\mu\left(x_n,t_0\right)$$

чтобы не вносить дополнительной погрешности $(\eta_1\equiv 0)$.

3.3 Устойчивость разностной схемы (14)

Напомним еще раз: *линейная схема (14) называется устойчивой по входным данным (по правой части и начальным условиям), если при достаточно малых h и τ существуют C_1,C_2 (не зависящие от h и τ), такие что,*

$$\|\delta y\|\leq C_1\|\delta\varphi\|+C_2\|\delta\chi\|,$$

то есть, решение непрерывно зависит от правой части и начальных условий.
Устойчивость разностной схемы, а следовательно и её сходимость при наличии аппроксимации, мы покажем в равномерной (чебышевской) метрике:

$$\|y\|_l=\max_{n,m}|y_n^m|$$

(сеточный аналог равномерной по t и x метрики).
Введем норму сеточного решения на m-ом слое:

$$\|y^m\|=\max_n|y_n^m|.$$

В силу Теоремы 1 (о достаточном условии равномерной устойчивости линейных разностных схем по начальным условиям) и Теоремы 2 (достаточного условия устойчивости линейной разностной схемы по правой части), нам достаточно показать, что, если существуют $C_1 \geq 0$ и $C_2 > 0$ и

$$\left\|y^{m+1}\right\| \leq (1 + \tau C_1) \left\|y^m\right\| + \tau C_2 \left\|\varphi\right\|, \qquad (*)$$

то схема устойчива по входным данным.
Ограничимся исследованием устойчивости в двух предельных случаях: чисто неявной ($\sigma = 1$) и чисто явной ($\sigma = 0$) схем.

а) Устойчивость чисто неявной схемы ($\sigma = 1$) : Рассмотрим разностное уравнениие (14.1):

$$\frac{1}{\tau}(\widehat{y} - y) = \alpha^2 \Lambda[\widehat{y}] + \varphi_n^m = \frac{a^2}{h^2}\left(y_{n+1}^{m+1} - 2y_n^{m+1} + y_{n-1}^{m+1}\right) + \varphi_n^m.$$

Обозначим $\gamma = \frac{\tau\alpha^2}{h^2}$, тогда

$$y_n^{m+1} - y_n^m = \gamma\left(y_{n+1}^{m+1} - 2y_n^{m+1} + y_{n-1}^{m+1}\right) + \tau\varphi_n^m \Longleftrightarrow$$

$$y_n^{m+1} = y_n^m - \gamma\left(2y_n^{m+1} - y_{n+1}^{m+1} - y_{n-1}^{m+1}\right) + \tau\varphi_n^m.$$

Покажем, что в этом случае ($\sigma = 1$) достаточное условие устойчивости ($*$) выполнено. Найдем на слое $(m+1)$ тот узел k_0, в котором y_n^{m+1} принимает наибольшее значение:

$$\max_n y_n^{m+1} = y_{k_0}^{m+1} \geq y_n^{m+1}, \quad \forall n.$$

Тогда

$$2y_{k_0}^{m+1} - y_{k_0+1}^{m+1} - y_{k_0-1}^{m+1} \geq 0.$$

Поэтому

$$y_{k_0}^{m+1} \leq y_{k_0}^m + \tau\varphi_{k_0}^m \leq \max_n y_n^m + \tau \max_{n,m} \varphi_n^m. \qquad (**)$$

С другой стороны, найдем на слое $(m+1)$ узел l_0 где y_n^{m+1} принимает минимальное значение:

$$\min_n y_n^{m+1} = y_{l_0}^{m+1} \leq y_n^{m+1}, \forall n.$$

Тогда

$$2y_{l_0}^{m+1} - y_{l_0+1}^{m+1} - y_{l_0-1}^{m+1} \leq 0$$

и

$$y_{l_0}^{m+1} \geq y_{l_0}^m + \tau\varphi_{l_0}^m \geq \min_n y_n^m + \tau \min_{n,m} \varphi_n^m. \qquad (***)$$

Объединяя $(**)$ и $(***)$, найдем:

$$\left\|y^{m+1}\right\| = \max_n |y_n^{m+1}| \leq \left\|y^m\right\| + \tau\left\|\varphi\right\|,$$

что совпадает с условием ($*$) при $C_1 = 0, C_2 = 1$. Таким образом, неявная схема ($\sigma = 1$) *безусловно* устойчива по входным данным (при любых τ и h).

б) Устойчивость чисто явной схемы ($\sigma = 0$) : Для чисто явной схемы уравнение (14.1) имеет вид:

$$\frac{1}{\tau}\left(y_n^{m+1} - y_n^m\right) = \alpha^2 \Lambda\left[y_n^m\right] + \varphi_n^m.$$

Откуда

$$y_n^{m+1} = y_n^m + \gamma\left(y_{n+1}^m - 2y_n^m + y_{n-1}^m\right) + \tau\varphi_n^m = (1-2\gamma)\, y_n^m + \gamma y_{n-1}^m + \gamma y_{n+1}^m + \tau\varphi_n^m.$$

Пусть $(1-2\gamma) > 0$, то есть $0 < \gamma < \frac{1}{2}$, тогда

$$\left|y_n^{m+1}\right| = \left|(1-2\gamma)\, y_n^m + \gamma y_{n-1}^m + \gamma y_{n+1}^m + \tau\varphi_n^m\right| \leq$$

$$\leq (1-2\gamma)\left|y_n^m\right| + \gamma\left|y_{n+1}^m\right| + \gamma\left|y_{n-1}^m\right| + \tau\left|\varphi_n^m\right|, \quad \forall n.$$

Тем самым

$$\|y^{m+1}\| \leq \|y^m\| + \tau\|\varphi\|, \quad C_1 = 0; C_2 = 1.$$

Итак, при

$$\gamma = \frac{\tau a^2}{h^2} < \frac{1}{2} \tag{15}$$

явная схема устойчива. Это условие накладывает жесткие ограничения на временной шаг сетки:

$$\tau < \frac{h^2}{2a^2}. \tag{15*}$$

Покажем, что при $\gamma > \frac{1}{2}$ явная схема *неустойчива* в чебышевской норме. Для этого достаточно показать, что, однажды возникнув, ошибка в решении будет при дальнейших вычислениях неограниченно возрастать. Рассмотрим однородную задачу (без правой части). Соответствующие возмущения — это возмущения начальных условий на данном слое. Схема при этом имеет вид

$$y_n^{m+1} = (1-2\gamma)\, y_n^m + \gamma y_{n-1}^m + \gamma y_{n+1}^m.$$

Пусть на m-ом слое возникла ошибка δy_n^m, тогда

$$\widetilde{y}_n^m = y_n^m + \delta y_n^m$$

и, поскольку $\widetilde{y}_n^m$ - это решение той же схемы,

$$\widetilde{y}_n^{m+1} = y_n^{m+1} + \delta y_n^{m+1} = (1-2\gamma)\left(y_n^m + \delta y_n^m\right) + \gamma\widetilde{y}_{n+1}^m + \gamma\widetilde{y}_{n-1}^m,$$

то в силу линейности нашей задачи, получаем уравнениие для ошибки:

$$\delta y_n^{m+1} = \delta y_n^m\,(1-2\gamma) + \gamma\delta y_{n+1}^m + \gamma\delta y_{n-1}^m.$$

Предположим, что ошибка является быстро осциллирующей функцией и имеет вид:

$$\delta y_n^m = (-1)^n\,\varepsilon; \quad \varepsilon > 0,$$

где ε - некоторое достаточно малое число, тогда:

$$\delta y_n^{m+1} = (1-2\gamma)(-1)^n\varepsilon + \gamma(-1)^{n+1}\varepsilon + \gamma(-1)^{n-1}\varepsilon = (-1)^n\varepsilon(1-4\gamma).$$

Но, так как $\gamma > 1/2$, то $4\gamma > 2$ и

$$\delta y_n^{m+1} = (-1)^{n+1}\varepsilon(4\gamma - 1).$$

Следовательно через k временных слоев

$$\left|\delta y_n^{m+k}\right| = \varepsilon\,(4\gamma - 1)^k \to \infty, \quad k \to \infty.$$

Уменьшение шага τ (при $\gamma > \frac{1}{2}$) не спасает, ибо при фиксированном T растет объем неустойчивых вычислений (за счет числа шагов), следовательно и ошибка. Значит явная схема $\sigma = 0$ при $\gamma = \frac{\tau a^2}{h^2} > \frac{1}{2}$ — неустойчива.

Замечания:

1) В силу устойчивости наших схем, мы показали, что $\|y^{m+1}\| \le \|y^m\| + \tau\,\|\varphi\|$. Это неравенство доказывает принцип максимума для наших схем: Пусть $\varphi = 0$ тогда

$$\left\|y^{m+1}\right\| \le \|y^m\| \le \ldots \le \left\|y^0\right\| = \|\chi\|$$

таким образом, во внутренних точках t и x норма решения не превосходит норму начальных условий.

2) В сеточном аналоге нормы L_2 методом гармоник (далее) можно показать, что схема (14) устойчива при

$$\sigma \ge \frac{1}{2} - \frac{h^2}{4\tau a^2}. \tag{15'}$$

В частности

a) $\sigma = \frac{1}{2}$ — безусловно устойчивая схема.

b) Схема с $\sigma = 0$ устойчива при условии

$$\frac{h^2}{4\tau a^2} \ge \frac{1}{2} \quad \Leftrightarrow \quad \frac{\tau a^2}{h^2} \le \frac{1}{-} \Leftrightarrow \quad \tau \le \frac{h^2}{2a^2}.$$

3) Можно показать, что в C схема (14) устойчива по входным данным при

$$\sigma \ge \frac{1}{2} - \frac{h^2}{2\tau a^2}. \tag{15''}$$

В частности схема с $\sigma = 0$ устойчива при условии

$$\tau \le \frac{h^2}{a^2}.$$

3.4 Сходимость разностной схемы (14)

Рассмотрим погрешность сеточного решения

$$z_n^m = y_n^m - u_n^m$$

$u_n^m = u\,(x_n, t_m)$ при простейшем способе проектирования $u(x,t)$ на сетку Ω.

Мы показали, что при наличии аппроксимации и устойчивости разностной схемы она обладает сходимостью, и порядок точности схемы (14) не ниже её порядка аппроксимации. В нашем случае имеет место равномерная сходимость либо сходимость

в среднем (в той же метрике, где есть и устойчивость). Поэтому для погрешности сеточного решения имеем оценки

a) $\sigma = \frac{1}{2}$:

$$\|z\|_c = O\left(h^2 + \tau^2\right)$$
$$u(x,t) \in C^{(4)}[0,l] \times C^{(3)}[0,T].$$

b) $\sigma \neq \frac{1}{2}$: (16)

$$\|z\|_c = O\left(\tau + h^2\right)$$
$$u(x,t) \in C^{(4)}[0,l] \times C^{(2)}[0,T]$$

При этом для обеспечения соответствующей аппроксимации, решение задачи (11) должно обладать указанной гладкостью.

3.5 Алгоритмы численного решения задачи (14). Прогонка

Сделаем краткое замечание относительно способов решения задачи (14).

а) В случае явной схемы ($\sigma = 0$). Алгоритм очевиден и определяется написанной явной формулой:

$$\begin{cases} y_n^{m+1} = (1-2\gamma)\, y_n^m + \gamma y_{n-1}^m + \gamma y_{n+1}^m + \tau\varphi_n^m; \quad 1 \le n \le N-1 \\ y_0^{m+1} = y_N^{m+1} = 0; \\ y_n^0 = \chi_n \end{cases} \tag{14*}$$

Напомним, что $\gamma < \frac{1}{2}$.

б) Для неявной схемы ($\sigma = 1$). Решение на $(m+1)$-ом временном слое находим из формул

$$\widehat{y}_n = y_n - \gamma\left(2\widehat{y}_n - \widehat{y}_{n+1} - \widehat{y}_{n-1}\right) + \tau\varphi_n^m,$$

что приводит к алгебраической системе

$$(1+2\gamma)\,\widehat{y}_n + \gamma\widehat{y}_{n-1} + \gamma\widehat{y}_{n+1} = y_n + \tau\varphi_n^m \Leftrightarrow$$
$$\begin{cases} A_n\widehat{y}_{n-1} + B_n\widehat{y}_n + C_n\widehat{y}_{n+1} = F_n; \\ \widehat{y}_0 = \widehat{y}_N = 0. \end{cases} \tag{14 * *}$$

Это СЛАУ с трехдиагональной матрицей, имеющей диагональное преобладание $B_n \ge A_n + C_n$ В таком случае решение $\widehat{y}_n$ существует и единственно. Решение дается формулами прогонки. Вычисления устойчивы. Общий объем вычислений при переходе на $(m+1)$-ый слой $O(9N)$ действий и требуется всего $O(3N)$ ячеек памяти для хранения матрицы СЛАУ.

Замечания: Мы рассмотрели однопараметрическое семейство схем (14) для одномерного уравнения теплопроводности.

Явная схема ($\sigma = 0$) алгоритмически наиболее проста, но требует выполнения жестких условий устойчивости $\tau < \frac{h^2}{2a^2}$, поэтому используется редко.

Широкое применение имеет схема $\sigma = \frac{1}{2}$, повышенной точности $O\left(h^2 + \tau^2\right)$ — безусловно устойчивая схема.

Схемы с $\sigma = \frac{1}{2}, \sigma = 1$ особенно эффективны для уравнениий с переменными коэффициентами или для квазилинейных уравнениий.

§4. Разностные схемы для одномерного уравнения колебаний

4.1 Постановка задачи. Разностная схема "крест"

Рассмотрим задачу для уравнения колебаний на отрезке с краевыми условиями 1-го рода (задачу Дирихле)

$$u_{tt} = a^2 u_{xx} + f(x,t), \qquad 0 < x < l, \quad t > 0$$

начальные условия

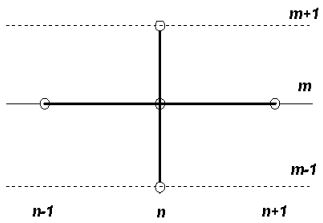
$$u(x,0) = \mu_1(x), \quad u_t(x,0) = \mu_2(x), \quad t = 0, \quad x \in [0,l] \tag{17}$$

краевые условия 1-го рода

$$u(0,t) = \mu_3(t) \equiv 0; \quad u(l,t) = \mu_4(t) \equiv 0, \quad t \geq 0$$

Введём обозначения

$$y = y_n^m; \quad \widehat{y} = y_n^{m+1}; \quad \check{y} = y_n^{m-1};$$



и, используя метод разностной аппроксимации, построим схему "крест" для одномерного уравнения теплопроводности.

Разностная аппроксимация самого уравнения:

$$\frac{1}{\tau^2}(\widehat{y} - 2y + \check{y}) = \frac{a^2}{h^2}(y_{n+1} - 2y_n + y_{n-1}) + \varphi_n^m. \tag{18.1}$$

При аппроксимации правой части мы использовали метод *неопределенных коэффициентов*.

Начальное условие для функции $u(x)$ аппроксимируется точно

$$y(x_n,0) = y_n^0 = \chi_{1n} = \mu_1(x_n) \Leftrightarrow \eta_1 \equiv 0.$$

Аппроксимация краевых условий также не вносит дополнительных погрешностей $\eta_3 \equiv 0$ и $\eta_4 \equiv 0$

$$y_0^m = \chi_3^m = \mu_3(t_m) = 0; \quad y_N^m = \chi_4^m = \mu_4(t_m) = 0.$$

При аппроксимации начального условия для производной

$$y_t(x_n,0) = \frac{y_n^1 - y_n^0}{\tau} = \chi_{2n}$$

порядок аппроксимации зависит от способа построения сеточной функции χ_2. Простейшая аппроксимация

$$\chi_{2n} = \mu_2(x_n) \quad \Rightarrow \quad \eta_2 \equiv O(\tau).$$

Если использовать само уравнение, то можно получить более аккуратную аппроксимацию начального условия

$$\frac{y_n^1 - y_n^0}{\tau} \Rightarrow \frac{u(x_n, \tau) - u(x_n, 0)}{\tau} = u'_t(x_n, 0) + \frac{\tau}{2} u''_{tt}(x_n, 0) + O(\tau^2)$$

Допустим:

$$u_{tt}(x_n, 0) = a^2 u_{xx}(x_n, 0) + f(x_n, 0) = a^2 \mu_{1xx}(x_n) + f_n^0.$$

При этом можно использовать аппроксимацию порядка $O(h^2)$ для $\mu_{1xx}(x_n)$. Таким образом

$$\frac{y_n^1 - y_n^0}{\tau} = \mu_2(x_n) + \frac{\tau}{2}\left(a^2 \mu_{1xx}(x_n, 0) + f_n^0\right); \quad \eta_2 = O(\tau^2).$$

Теперь запишем разностную схему для исходной задачи (17)

$$\frac{\widehat{y} - 2y + \check{y}}{\tau^2} = a^2 \, \frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} + \varphi_n^m$$

краевые условия

$$y_0^m = \chi_3^m = \mu_3(t_m) \equiv 0; \quad y_n^m = \chi_4^m = \mu_4(t_m) \equiv 0 \tag{18}$$

начальные условия

$$y_n^0 = \chi_1(x_n) = \mu_1(x_n)$$

$$\frac{y_n^1 - y_n^0}{\tau} = \chi_2(x_n) = \begin{cases} \mu_2(x_n) & \Rightarrow \quad \eta_2 = O(\tau) \\ \mu_2(x_n) + \dfrac{\tau}{2}\left(a^2 \mu_{1xx}(x_n) + f(x_n, 0)\right) & \Rightarrow \quad \eta_2 = O(\tau^2). \end{cases}$$

Это *явная схема* относительно $\widehat{y} \equiv y_n^{m+1}$. После того, как найдено $\{y_n^1\}$ из начального условия далее расчётные формулы просты.

Задача 1°: *Получить расчетные формулы для* $\widehat{y}$.

4.2 Порядок аппроксимации разностной схемы (18)

Сам принцип построения разностной схемы (18) позволяет утверждать, что:

1) $\varphi_n^m = f(x_n, t_m)$ — необходимое условие для аппроксимации;

2) порядок аппроксимации (18.1) есть $O(\tau^2 + h^2)$ в силу симметрии полученных разностных формул;

3) с учетом (18.2) $\Rightarrow$ общий порядок аппроксимации схемы $O(\tau^2 + h^2)$.

Задача 2°: *Подтвердить перечисленные пункты.*

4.3 Устойчивость разностной схемы (18)

Для доказательства устойчивости схемы (18) используем *метод разделения переменных* (поскольку коэффициенты схемы постоянны или их можно "заморозить" на данном временном слое) или *метод гармоник*. Этим методом доказывается устойчивость разностной схемы в сеточном аналоге $\mathcal{L}_2$ ("в среднем").

На каждом временном слое сеточная функция по $\{x_n\}$ может быть разложена по собственным сеточным функциям сеточного оператора Лапласа $\Lambda_{\overline{x}x}$ это "косинусы" и "синусы" от $\left(\frac{\pi k}{l}x_n\right)$ для k-ой функции. Поведение гармоник на различных слоях по t характеризуется множителями роста гармоники ρ_k, т.е. рассматривается устойчивость решения вида

$$y_n^m = y\left(x_n, t_m\right) = \left(\rho_k\right)^m e^{ikx_n}, \qquad k = 0; \pm 1; \pm 2; ...$$

Имеет место теорема:

Теорема 5. *Двуслойная разностная схема с постоянными коэффициентами устойчива в среднем по начальным данным , если $\forall$ k (т.е. для любой гармоники) множитель роста удовлетворяет условию*

$$|\rho_k| \leq 1 + C\tau; \qquad C \geq 0 \quad - const. \tag{$*$}$$

Ограничимся замечаниями:

1) Фактически $const$ $C \geq 0$ не должна быть очень большой. На практике условие (*) проверяют для $C = 0$, т.е. $|\rho_k| \leq 1$.

2) Условие (*) в некотором смысле и необходимо, т.е. если существует гармоника k_0 для которй (*) не выполняется, то схема неустойчива.

Теперь вернемся к нашей задаче (18). Пусть $\varphi_n^m \equiv 0$; $y_n^m = e^{ikx_n}$ — начнем с этого слоя. Тогда

$$\widehat{y} = \rho_k e^{ikx_n} = \rho_k y; \quad \check{y} \Rightarrow y = \rho_k \check{y}.$$

Однородное уравнение (18.1) даёт

$$\left(\rho_k - 2 + \frac{1}{\rho_k}\right) = \underbrace{\frac{\tau^2 a^2}{h^2}}_{\gamma^2}\left(e^{ikh} - 2 + e^{-ikh}\right) = \gamma^2\left(2\cos kh - 2\right) = -4\gamma^2 \sin^2\frac{kh}{2}.$$

Множители роста k-ой гармоники ρ_k удовлетворяют уравнению

$$\rho_k^2 + 2\rho_k\left(1 - 2\gamma^2 \sin^2\frac{kh}{2}\right) + 1 = 0. \tag{$**$}$$

По теореме Виетта — $\left(\rho_k\right)_1\left(\rho_k\right)_2 = 1$ и требование устойчивости $\left|\left(\rho_k\right)_{1,2}\right| \leq 1$ выполнено, если только

$$\left|\left(\rho_k\right)_{1,2}\right| = 1.$$

Значит $(\rho_k)_1$ и $(\rho_k)_2$ – комплексно-сопряженные числа. Это в свою очередь возможно лишь при отрицательном дискриминанте уравнения (**) $D < 0$.

Итак

$$\left(1 - 2\gamma^2 \sin^2\frac{kh}{2}\right)^2 - 1 < 0 \quad \Leftrightarrow \quad \left|1 - 2\gamma^2 \sin^2\frac{kh}{2}\right| < 1.$$

Это условие относительно γ (точнее τ и h) и оно заведомо верно $\forall k$, если $\gamma^2 < 1$, т.е.

$$\frac{\tau a}{h} < 1 \quad \text{— условие Куранта.} \tag{19}$$

Замечания:
1) Схема "крест" устойчива в среднем по начальным данным при дополнительном условии $\tau a/h < 1$.
2) При условии (19) схема "крест" устойчива по правой части;
3) При условии (19) схема "крест" устойчива по начальным данным и правой части в равномерной сеточной норме (в C).

4.4 Сходимость схемы "крест"

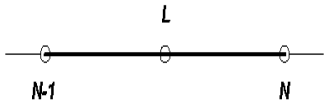
Установленный нами порядок аппроксимации и устойчивость схемы (18) позволяет утверждать наличие сходимости схемы (в соответствующей метрике) с точностью не ниже порядка аппроксимации. Итак

$$\|z\| = \|y - u\| = \begin{cases} O\left(\tau + h^2\right) \\ O\left(\tau^2 + h^2\right) \end{cases} \quad \text{при} \quad \frac{\tau a}{h} < 1 \Leftrightarrow \tau < \frac{h}{a}.$$

Сходимость указанных порядков возможна лишь для решений, обладающих достаточной гладкостью, чтобы обеспечить аппроксимацию этих порядков.
Достаточно

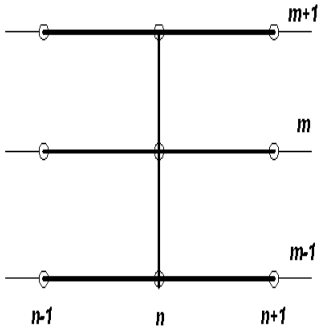
$$u(x,t) \in C^{(4)}[0,l] \times C^{(4)}[0,T].$$

Замечания к § 4:
1) При аппроксимации краевых условий 2-го рода, например $u_x(l,t) = \mu_4(t)$, удобно сетку по x строить так, чтобы точка $x = l$ оказалась бы между узлами сетки, тогда



$$\frac{y_N^m - y_{N-1}^m}{h} = \mu_4(t_m) \quad \Rightarrow \quad \eta_4 = O\left(h^2\right).$$

2) Не представляет труда построить для одномерного уравнеия колебаний неявную 9-ти точечную схему с весами.



В шаблоне использованы три временных слоя. Основное уравнение схемы

$$\frac{1}{\tau^2}\left(\widehat{y} - 2y + \check{y}\right) = a^2 \Lambda_{\bar{x}x}\left[\sigma\widehat{y} + (1 - 2\sigma)\, y + \sigma\check{y}\right] + \varphi_n^m,$$

где $0 \le \sigma \le \frac{1}{2}$.

Задача 3.
1) Установить порядок аппроксимации такой схемы.
2) Показать, что при $\frac{1}{4} \le \sigma \le \frac{1}{2}$ схема безусловно устойчива в $\mathcal{L}_2$ по начальным данным.

§5. Многомерные разностные схемы для уравнения теплопроводности

Рассмотрим задачу о распределении тепла в прямоугольной области:

$$\begin{cases} u_t = a^2\left(u_{x_1x_1} + u_{x_2x_2}\right) + f\left(x_1, x_2, t\right), & \begin{matrix} 0 < x_1 < l_1 \\ 0 < x_2 < l_2 \\ 0 < t \end{matrix} \\ u\,|_\Gamma = \mu_\Gamma\left(t\right) \text{ (задача Дирихле)} \\ u\,|_{t=0} = \mu\left(x_1, x_2\right) \end{cases} \tag{20}$$

Будем предполагать, что задача (20) корректна и входные данные обеспечивают нужную гладкость решения.

5.1 Разностная схема

Обобщим на задачу (20) схемы § 3. Рассмотрим в $\overline{\mathcal{D}}$ равномерную сетку:

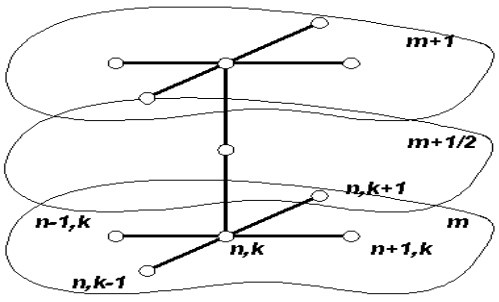
$$\overline{\omega}_{h_1,h_2,\tau} = \left\{ \left(x_{1n}, x_{2k}, t_m\right): \begin{matrix} x_{1n} = nh_1;\ n = \overline{1,N};\ h_1 = \dfrac{l_1}{N} \\ x_{2k} = kh_2;\ k = \overline{1,K};\ h_2 = \dfrac{l_2}{K} \\ t_m = m\tau;\ m = \overline{1,M};\ \tau = \dfrac{T}{M} \end{matrix} \right\}.$$

Граничные условия аппроксимируются в этом случае точно:

$$\chi_\Gamma^m = \mu_\Gamma\left(t_m\right); \quad \eta_\Gamma = 0; \quad \begin{matrix}\text{на каждой} \\ \text{стороне прямоуголь-} \\ \text{ника,}\end{matrix}$$

поскольку точки сетки естественным образом задают границу области $\overline{\mathcal{D}}$. (Дискуссию по этому вопросу мы исключим).

Пусть $y_{n,k}^m = y$; $\widehat{y} = y_{n,k}^{m+1}$. Составим двуслойную схему с весами. Аппроксимируем оператор Лапласа Δ_2 $\Rightarrow$



$$\Lambda = \Lambda_1 + \Lambda_2, \qquad \text{где}$$

$$\Lambda_1 y = \Lambda_{\overline{x}_1 x_1} y = \frac{1}{h_1^2}\left(y_{n+1,k} - 2y_{n,k} + y_{n-1,k}\right)$$

$$\Lambda_2 y = \Lambda_{\overline{x}_2 x_2} y = \frac{1}{h_2^2}\left(y_{n,k+1} - 2y_{n,k} + y_{n,k-1}\right).$$

Эти операторы аппроксимируют $\dfrac{\partial^2}{\partial x_1^2}$ и $\dfrac{\partial^2}{\partial x_2^2}$ со вторым порядком по пространственным переменным. Сеточный оператор $(\Lambda_1 + \Lambda_2)$ аппроксимирует оператор Лапласа

Δu в узле (n,k) с невязкой $O\left(h_1^2+h_2^2\right)$.
Тогда основное уравнение задачи (20) аппроксимируется разностным уравнением

$$\frac{1}{\tau}\left(\widehat{y}_{n,k}-y_{n,k}\right)=a^2\left(\Lambda_1+\Lambda_2\right)\left[\sigma\widehat{y}_{n,k}+(1-\sigma)\,y_{n,k}\right]+\varphi_{n,k}^m \tag{21.1}$$

Задача 4.

1) Установить порядок аппроксимации для уравнения (21.1)

2) Показать, что $\varphi_{n,k}^m=f\left(x_{1n},x_{2k},t_{m+\frac{1}{2}}\right)=\overline{f}_{n,k}$.

3) Показать, что

при $\sigma=\frac{1}{2}$ невязка $\psi=O\left(\tau^2+h_1^2+h_2^2\right)$

при $\sigma\neq\frac{1}{2}$ невязка $\psi=O\left(\tau+h_1^2+h_2^2\right)$.

Задача 5.

1)Методом разделения переменных доказать устойчивость в $\mathcal{L}_2$ схемы (21) по начальным условиям при

$$\sigma\geq\frac{1}{2}-\frac{\left(\frac{1}{h_1^2}+\frac{1}{h_2^2}\right)^{-1}}{4\tau a^2}. \tag{22}$$

2)Получить условие устойчивости для явной схемы $\sigma=0$.

Существенный недостаток схемы (21) в *многомерном* случае связан с тем, что как чисто явная схема $\sigma=0$, так и неявная $\sigma\neq0$ 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 $\sigma=0$ схемы для построения решения на последнем слое T есть $O(N^4)$. Действительно, для перехода на следующий временной слой решается явная система уравнений с числом неизвестных $O(NK)\sim O(N^2)$. 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 $O(N^4)$.

Для неявный схем $\sigma\neq0$ положение ещё чуже. Ограничиваясь абсолютно устойчивым вариантом схем при $\sigma\geq\frac{1}{2}$, на каждом временном слое приходится решать СЛАУ с N^2 уравнений при ширине ленты порядка $O(2N)$. Метод исключения Гаусса требует $O(N^6)$ с учётом ленточной структуры матрицы — $O(N^4)$ действий. Требование аппроксимации даёт $O(N)$ шагов по времени. Итого — $O(N^5)$! Неявная схема менее выгодна в этом случае!

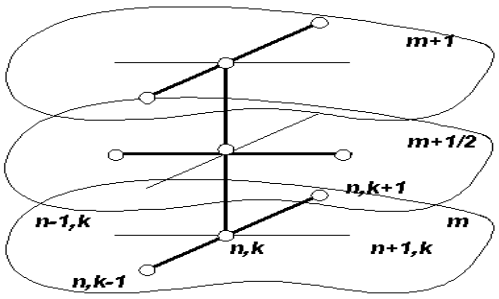
Поэтому предпочтение отдают абсолютно устойчивым ($\tau\sim h$), экономичным разностным схемам, в которых при переходе на очередной временной слой совершается всего $O(N^2)$ действий.

§6. Продольно-поперечная разностная схема для уравнения теплопроводности. Экономичные разностные схемы

Введем промежуточный по t слой $(m+\frac{1}{2})$ и рассмотрим разностную схему

$$\frac{\overline{y}_{n,k}-y_{n,k}}{\tau/2}=a^2\Lambda_1\overline{y}_{n,k}+a^2\Lambda_2 y_{n,k}+\overline{f}_{n,k}$$

$$\frac{\widehat{y}_{n,k}-\overline{y}_{n,k}}{\tau/2}=a^2\Lambda_1\overline{y}+a^2\Lambda_2\widehat{y}+\overline{f}_{n,k}. \qquad (23)$$



Обсудим построение решения уравнения (23) на $(m+1)$ слое:

1)Уравнение (23.1) позволяет найти $\overline{y}_{n,k}$ по неявной схеме относительно x_1 и по явной схеме относительно x_2 $\Rightarrow$ Решается система с 3-х диагональная матрицей относительно переменной x_1 эффективным методом прогонки по x_1 при каждом k (k - раз прогонка с $O(N)$ действий $\Rightarrow$ $O(NK)$ действий).

2)Уравнение (23.2) позволяет найти $\widehat{y}_{n,k}$ по неявной схеме относительно x_2 и по явной схеме относительно x_1 $\Rightarrow$ прогонка по x_2 при каждом n $\Rightarrow$ $O(N\,K)$ действии $\Rightarrow$ итого $O(2N\,K)\sim O(N^2)$ 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 $O(30N^2)$ действий.

Другие достоинства схемы:

6.1 Устойчивость продольно-поперечной схемы

Воспользуемся методом гармоник. Рассмотрим

$$y_{n,k}=\exp\left(ix_{1n}p+ix_{2k}q\right);\quad \overline{y}=p'_{p,q}\,y;\quad \widehat{y}=p''_{p,q}\,\overline{y}$$

(свои множители роста на каждом полуслое). Тогда (23.1):

$$p'_{p,q}-1=\frac{\tau a^2}{2h_1^2}\left(-4\sin^2\frac{ph_1}{2}\right)p'_{p,q}+\frac{\tau a^2}{2h_2^2}\left(-4\sin^2\frac{qh_2}{2}\right),\qquad \text{т.е.}$$

$$\text{Аналогично}\quad \left.\begin{aligned} p'_{p,q} &= \frac{1-\frac{2\tau a^2}{h_2^2}\sin^2\frac{qh_2}{2}}{1+\frac{2\tau a^2}{h_1^2}\sin^2\frac{ph_1}{2}} \\ p''_{p,q} &= \frac{1-\frac{2\tau a^2}{h_1^2}\sin^2\frac{ph_1}{2}}{1+\frac{2\tau a^2}{h_2^2}\sin^2\frac{qh_2}{2}} \end{aligned}\right\} \quad \Rightarrow \quad |p_{pq}| = \left|p'_{p,q}\,p''_{p,q}\right| \le 1$$

всегда! $\forall\, p$ и q. Таким образом схема (23) безусловно (абсолютно) устойчивагв $\mathcal{L}_2$ по начальным данным (и по правой части тоже).
Для рассмотренной схемы имеет место абсолютная устойчивость в C по начальным условиям и по правой части.

Осталось установить аппроксимацию.

6.2 Аппроксимация продольно-поперечной схемы

Исключим из (23) слой $\overline{y}_{n,k}$. Для этого вычтем уравнения (1)-(2), найдём:

$$2\frac{\overline{y}_{nk}}{\tau/2} - \frac{\widehat{y}_{nk}+y_{nk}}{\tau/2} = -a^2\Lambda_2\left(\widehat{y}-y\right), \qquad \text{т.е.}$$

$$\overline{y}_{nk} = \frac{\widehat{y}_{nk}+y}{2} - \frac{\tau a^2}{4}\Lambda_2\left(\widehat{y}-y\right); \qquad (*)$$

Складывая уравнения (1) - (2), найдем:

$$\frac{\widehat{y}_{nk}-y_{nk}}{\tau/2} = a^2\Lambda_1\left(2\overline{y}_{nk}\right) + a^2\Lambda_2\left(\widehat{y}_{nk}+y_{nk}\right) + 2\overline{f}_{nk}.$$

Откуда, с учетом (*), получим

$$\frac{\widehat{y}_{nk}-y_{nk}}{\tau} = a^2\Lambda_1\left(\frac{\widehat{y}_{nk}+y}{2}\right) - \frac{\tau a^2}{4}\Lambda_1\Lambda_2\left(\widehat{y}-y\right) + a^2\Lambda_2\frac{\widehat{y}_{nk}+y_{nk}}{2} + \overline{f}_{nk} =$$

$$= a^2\left(\Lambda_1+\Lambda_2\right)\frac{\widehat{y}_{nk}+y_{nk}}{2} - \underbrace{\frac{\tau a^2}{4}\Lambda_1\Lambda_2\overbrace{\left(\widehat{y}-y\right)}^{O(\tau)}}_{O(\tau^2)} + \overline{f}_{nk}.$$

Итак, это почти симметричная схема с $\sigma_1 = \sigma_2 = \frac{1}{2}$, тем самым — схема обладает аппроксимацией при условии $\overline{f}_{nk} = f\left(x_{1n}, x_{2n}, t_{m+\frac{1}{2}}\right)$ и порядок аппроксимации

$$\psi = O\left(\tau^2 + h_1^2 + h_2^2\right).$$

Схема (23) безусловно устойчива и обладает повышенной аппроксимацией следовательно она сходится в указанной прямоугольной области на равномерной сетке и обладает точностью не хуже, чем

$$\|y-u\| = O\left(\tau^2 + h_1^2 + h_2^2\right).$$

Замечания:

1) Схема обладает той же сходимостью в C.

2) Для обеспечения указанного порядка точности разностной схемы требуется, чтобы решения исходной задачи обладали гладкостью не хуже, чем

$$u(x_1, x_2, t) \in C^{(5)}([0; l_1] \times [0; l_2]) \bigcup C^{(5)}[0; T].$$

Вместо заключения

Рассмотренный нами лекционный материал исчерпывает предложенную Вашему вниманию программу лекционного курса. Хотя при этом мы лишь подошли ко многим наиболее интересным и содержательным с прикладной точки зрения вопросам. Но, позволю себе напомнить, что "... достаточно иллюзорной представляется попытка изложить "во всей полноте и строгости" курс численных методов, поэтому основная цель лекционного курса преследовала задачу собрать и изложить минимальный материал достаточный для дальнейшей самостоятельной работы в области разумного применения и создания новых численных методов."

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 1.**

1. Решение задачи аппроксимации методом наименьших квадратов.
2. Разностные схемы для уравнения теплопроводности в прямоугольной области. Экономичные разностные схемы.
3. Формула погрешности $\delta\lambda$ собственного значения симметричной квадратной матрицы.

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 2.**

1. Погрешность округления на t-разрядной ЭВМ.
2. Разностная схема для одномерного волнового уравнения. Постановка задачи. Аппроксимация начальных условий.
3. Формула погрешности δx собственного вектора симметричной квадратной матрицы

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 3.**

1. Постановка задачи интерполяции. Чебышевская система интерполяционных функций.
2. Порядок точности линейной разностной схемы.
3. Получить оценку остаточного члена формулы численного интегрирования трапеций.

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 4.**

1. Полиномиальная интерполяция.
2. Сходимость разностной схемы.
3. Получить оценку остаточного члена формулы численного интегрирования Симпсона.

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 6.**

1. Интерполяционный многочлен Ньютона.
2. Продольно-поперечная разностная схема для уравнения теплопроводности. Устойчивость.
3. Получить оценку остаточного члена формулы численного интегрирования трапеций.

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 5.**

1. Погрешность полиномиальной интерполяции на равномерной сетке.
2. Метод Ньютона-Рафсона минимизации функции многих переменных.
3. Исследовать невязку схемы "крест" (записать недостающие уравнения)

$$\frac{1}{\tau^2}(\hat{y}_n - 2y_n + \check{y}_n) = \frac{a^2}{h^2}(y_{n+1} - 2y_n + y_{n-1}) + f_n, \quad 1 \le n \le N-1.$$

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 7.**

1. Устоичивость задачи определения собственных значений и собственных векторов квадратной матрицы.
2. Разностная схема для одномерного уравнения колебаний. Постановка задачи. Аппроксимация начальных условий.
3. Формулы прогонки решения СЛАУ с трехдиагональной матирцей.

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 8.**

1. Существование и единственность наилучшего среднеквадратичного приближения функции.
2. Свойства сопряжённых направлений в задаче минимизации функции многих переменных.
3. Исследовать устойчивость схемы "крест" (записать недостающие уравнения)

$$\frac{1}{\tau^2}(\hat{y}_n - 2y_n + \check{y}_n) = \frac{a^2}{h^2}(y_{n+1} - 2y_n + y_{n-1}) + f_n, \quad 1 \le n \le N-1.$$

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 9.**

1. Погрешность округления на t-разрядной ЭВМ.
2. Итерационные методы решения СЛАУ. Метод Зейделя, метод верхней релаксации.
3. Исследовать невязку схемы с весами (записать недостающие уравнения)

$$\frac{1}{\tau^2}(\hat{y}_n - 2y_n + \check{y}_n) = a^2\Lambda[\sigma\hat{y} + (1-2\sigma)y + \sigma\check{y}] + f_n, \quad 0 \le \sigma \le 1/2.$$

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 10.**

1. Пстроение кубического интерполяционного сплайна.
2. Учёт погрешностей округления при решении систем линейных алгебраических уравнений.
3. Исследовать устойчивость схемы с весами (записать недостающие уравнения)

$$\frac{1}{\tau^2}(\hat{y}_n - 2y_n + \check{y}_n) = a^2\Lambda[\sigma\hat{y} + (1-2\sigma)y + \sigma\check{y}] + f_n, \quad 1/4 \le \sigma \le 1/2.$$

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 11.**

1. Существование и единственность интерполяционного кубического сплайна.
2. Итерационные методы решения СЛАУ. Метод релаксации, метод Якоби.
3. Исследовать устойчивость схемы с весами (записать недостающие уравнения)

$$\frac{1}{\tau^2}(\hat{y}_n - 2y_n + \check{y}_n) = a^2\Lambda[\sigma\hat{y} + (1-2\sigma)y + \sigma\check{y}] + f_n, \quad 0 \le \sigma < 1/4.$$

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 12.**

1. Постановка задачи аппроксимации функции.
2. Продольно-поперечная схема для уравнения теплопроводности. Аппроксимация.
3. Получить оценку остаточного члена формулы численного интегрирования трапеций.

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 13.**

1. Существование и единственность наилучшего среднеквадратичного приближения функции.
2. Минимум функции многих переменных. Квадратичная функция, её свойста, экстремумы.
3. Исследовать невязку схемы с весами (записать недостающие уравнения)

$$\frac{1}{\tau^2}(\hat{y}_n - 2y_n + \check{y}_n) = a^2\Lambda[\sigma\hat{y} + (1-\sigma)\check{y}] + f_n, \quad 0 \le \sigma \le 1.$$

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 14.**

1. Ортогональные системы полиномов: Якоби, Лежандра, Чебышева 1-го и 2-го рода, Лагерра, Эрмита.
2. Устойчивость решения СЛАУ.
3. Исследовать устойчивость схемы с весами (записать недостающие уравнения)

$$\frac{1}{\tau^2}(\hat{y}_n - 2y_n + \check{y}_n) = a^2\Lambda[\sigma\hat{y} + (1-\sigma)\check{y}] + f_n, \quad 0 \le \sigma < 1/2.$$

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 15.**

1. Решение задачи аппроксимации методом наименьших квадратов.
2. Формулы матричной прогонки.
3. Исследовать устойчивость схемы с весами (записать недостающие уравнения)

$$\frac{1}{\tau^2}(\hat{y}_n - 2y_n + \check{y}_n) = a^2\Lambda[\sigma\hat{y} + (1-\sigma)\check{y}] + f_n, \quad 0 \le 1/2 \le \sigma < 1.$$

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 16.**

1. Аппроксимация 2π-периодической функции на равномерной сетке. Формулы Бесселя.
2. Нахождение собственных векторов квадратной матрицы методом обратной итерации.
3. Исследовать аппроксимацию схемы (одномерное уравнение теплопроводности)

$$\frac{1}{2\tau}(\hat{y}_n - \check{y}_n) = \frac{a^2}{h^2}(y_{n+1} - 2y_n + y_{n-1}) + f_n.$$

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 17.**

1. Сглаживание таблиц методом наименьших квадратов.
2. Нахождение собственных значений квадратной матрицы методом интерполяции. Случай 3-х диагональной матрицы.
3. Показать, что схема (одномерное уравнение теплопроводности)
$$\frac{1}{2\tau}(\hat{y}_n - \check{y}_n) = \frac{a^2}{h^2}(y_{n+1} - 2y_n + y_{n-1}) + f_n$$
абсолютно неустойчива.

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 18.**

1. Постановка задачи численного интегрирования. Квадратурные формулы интерполяционного типа.
2. Устойчивость задачи на собственные значения для эрмитовских матриц.
3. Исследовать аппроксимацию схемы (одномерное уравнение теплопроводности)
$$\frac{1}{2\tau}(\hat{y}_n - \check{y}_n) = \frac{a^2}{h^2}(y_{n+1} - 2 \cdot \frac{\hat{y}_n + \check{y}_n}{2} + y_{n-1}) + f_n.$$

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 19.**

1. Квадратурные формулы Ньтона-Котесса.
2. Вычисление определителя. Построение обратной матрицы.
3. Исследовать устойчивость схемы (одномерное уравнение теплопроводности)
$$\frac{1}{2\tau}(\hat{y}_n - \check{y}_n) = \frac{a^2}{h^2}(y_{n+1} - 2 \cdot \frac{\hat{y}_n + \check{y}_n}{2} + y_{n-1}) + f_n.$$

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 20.**

1. Квадратурная формула трапеций. Оценка остаточного члена квадратурной формулы.
2. Метод "золотого сечения" поиска минимума функции одного переменного.
3. Исследовать устойчивость схемы с весами (записать недостающие уравнения)
$$\frac{1}{\tau^2}(\hat{y}_n - 2y_n + \check{y}_n) = a^2\Lambda[\sigma\hat{y} + (1-\sigma)\check{y}] + f_n, \quad 0 \le 1/2 \le \sigma < 1.$$

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 21.**

1. Квадратурная формула Симпсона. Оценка остаточного члена квадраткрной формулы.

2. Метод "парабол" поиска минимума функции одного переменного.

3. Исследовать аппроксимацию схемы (одномерное уравнение теплопроводности)

$$\frac{1}{2\tau}(\hat{y}_n - \check{y}_n) = \frac{a^2}{h^2}(y_{n+1} - 2y_n + y_{n-1}) + f_n.$$

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 22.**

1. Апостериорная оценка погрешности квадратурных формул. Метод Рунге.

2. Сходимость итерационных методов решения СЛАУ. Достаточные условия.

3. Исследовать устойчивость схемы с весами (записать недостающие уравнения)

$$\frac{1}{\tau^2}(\hat{y}_n - 2y_n + \check{y}_n) = a^2\Lambda[\sigma\hat{y} + (1-\sigma)\check{y}] + f_n, \quad 0 \le \sigma < 1/2.$$

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 23.**

1. Вывод составной квадратурной формулы Симпсона. Погрешность.

2. Достаточное условие сходимости итерационных методов решения СЛАУ для симметричной, положительно определенной матрицы.

3. Показать, что схема (одномерное уравнение теплопроводности)

$$\frac{1}{2\tau}(\hat{y}_n - \check{y}_n) = \frac{a^2}{h^2}(y_{n+1} - 2y_n + y_{n-1}) + f_n$$

абсолютно неустойчива.

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 24.**

1. Вывод составной формулы трапеций. Погрешность.

2. LU-разложение невырожденной матрицы.

3. Исследовать устойчивость схемы с весами (записать недостающие уравнения)

$$\frac{1}{\tau^2}(\hat{y}_n - 2y_n + \check{y}_n) = a^2\Lambda[\sigma\hat{y} + (1-2\sigma)y + \sigma\check{y}] + f_n, \quad 0 \le \sigma < 1/4.$$

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 26.**

1. Квадратурные формулы Гаусса-Кристоффеля.
2. Невязка разностной схемы.
3. Формула погрешности δx собственного вектора симметричной квадратной матрицы

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 27.**

1. Корректность формул численного интегрирования.
2. Метод сопряжённых градиентов минимизации функции многих переменных.
3. Формула погрешности $\delta\lambda$ собственного значения симметричной квадратной матрицы.

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 25.**

1. Апостериорная оценка погрешности квадратурной формулы. Метод Эйткена.
2. Обусловленность матрицы СЛАУ. Относительнач погрешность решения.
3. Исследовать аппроксимацию схемы (одномерное уравнение теплопроводности)

$$\frac{1}{2\tau}(\hat{y}_n - \check{y}_n) = \frac{a^2}{h^2}(y_{n+1} - 2 \cdot \frac{\hat{y}_n + \check{y}_n}{2} + y_{n-1}) + f_n.$$

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 28.**

1. Интегрирование быстро осциллирующих функций методом Филона.
2. Характерный вид рельефа поверхности уровня в окрестности точки возможного экстремума. Котловинный и овражный рельефы.
3. Исследовать устойчивость схемы (одномерное уравнение теплопроводности)

$$\frac{1}{2\tau}(\hat{y}_n - \check{y}_n) = \frac{a^2}{h^2}(y_{n+1} - 2 \cdot \frac{\hat{y}_n + \check{y}_n}{2} + y_{n-1}) + f_n.$$

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 29.**

1. Постановка задачи решения нелинейного уравнения. Метод простой итерации.
2. Разностная схема для уравнения теплопроводности на отрезке. Устойчивость явной разностной схемы.
3. Получить оценку остаточного члена формулы численного интегрирования трапеций.

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 30.**

1. Сходимость метода простой итерации. Принцип сжимающих отображений.
2. Метод прогонки решения СЛАУ с $3\overline{-x}$-диагональной матрицей.
3. Исследовать невязку схемы "крест" (записать недостающие уравнения)

$$\frac{1}{\tau^2}(\hat{y}_n - 2y_n + \check{y}_n) = \frac{a^2}{h^2}(y_{n+1} - 2y_n + y_{n-1}) + f_n, \quad 1 \le n \le N-1.$$

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 31.**

1. Необходимые и достаточные условия существования решения нелинейного уравнения. Примеры итерационных методов решения $f(x) = 0$.
2. Метод последовательного исключения Гаусса решения СЛАУ.
3. Исследовать устойчивость схемы с весами (записать недостающие уравнения)

$$\frac{1}{\tau^2}(\hat{y}_n - 2y_n + \check{y}_n) = a^2\Lambda[\sigma\hat{y} + (1-\sigma)\check{y}] + f_n, \quad 0 \le \sigma < 1/2.$$

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 32.**

1. Ортогональные системы полиномов: Якоби, Лежандра, Чебышева 1-го и 2-го рода, Лагерра, Эрмита.
2. Нахождение собственных векторов квадратной матрицы методом обратной итерации.
3. Показать, что схема (одномерное уравнение теплопроводности)

$$\frac{1}{2\tau}(\hat{y}_n - \check{y}_n) = \frac{a^2}{h^2}(y_{n+1} - 2y_n + y_{n-1}) + f_n$$

абсолютно неустойчива.

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 33.**

1. Сходимость итерационных методов решения нелинейного уравнения. Достаточные условия.
2. Постановка задачи минимизации функционала. Метод пробных функций.
3. Исследовать устойчивость схемы с весами (записать недостающие уравнения)

$$\frac{1}{\tau^2}(\hat{y}_n - 2y_n + \check{y}_n) = a^2\Lambda[\sigma\hat{y} + (1-2\sigma)y + \sigma\check{y}] + f_n, \quad 0 \le \sigma < 1/4.$$

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 34.**

1. Итерационные методы p-го порядка. Примеры для $p = 2$.
2. Метод вращений Якоби нахождения собственных значений и собственных векторов действительной симметричной матрицы.
3. Исследовать невязку схемы "крест" (записать недостающие уравнения)

$$\frac{1}{\tau^2}(\hat{y}_n - 2y_n + \check{y}_n) = \frac{a^2}{h^2}(y_{n+1} - 2y_n + y_{n-1}) + f_n, \quad 1 \le n \le N-1.$$

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 35.**

1. Достаточные условия сходимости метода релаксации решения нелинейного уравнения.
2. Устойчивость разностной схемы.
3. Формула погрешности δx собственного вектора симметричной квадратной матрицы

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 36.**

1. Достаточные условия сходимости метода Ньютона решения нелинейного уравнения.
2. Порядок точности разностной схемы с весами для одномерного уравнения теплопроводности на отрезке.
3. Формула погрешности $\delta\lambda$ собственного значения симметричной квадратной матрицы.

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 37.**

1. Ускорение сходимости итерационных методов первого порядка.
2. Порядок аппроксимации разностной схемы.
3. Получить оценку остаточного члена формулы численного интегрирования трапеций.

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 38.**

1. Итерационные методы решения систем нелинейных уравнений. Одношаговые итерационные методы.
2. Устойчивость двуслойных разностных схем по начальным данным.
3. Формулы прогонки решения СЛАУ с трехдиагональной матрицей.

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 39.**

1. Сходимость метода Ньютона решения системы нелинейных уравнений.
2. Устойчивость двуслойной разностной схемы по правым частям.
3. Получить оценку остаточного члена формулы численного интегрирования трапеций.

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 40.**

1. Достаточные условия сходимости метода покоординатного спуска минимизации функции многих переменных.
2. Разностная схема с весами для одномерного уравнения теплопроводности на отрезке. Порядок аппроксимации.
3. Формулы прогонки решения СЛАУ с трехдиагональной матрицей.

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 41.**

1. Погрешность решения: погрешность модели, неустранимая погрешность, погрешность метода, погрешности округления.
2. Разностная схема для уравнения теплопроводности на отрезке. Устойчивость неявной схемы.
3. Получить оценку остаточного члена формулы численного интегрирования Симпсона.

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 42.**

1. Постановка задачи интерполяции. Чебышевская система интерполяционных функций.
2. Разностная схема для одномерного волнового уравнения на отрезке. Порядок аппроксимации.
3. Формула погрешности δx собственного вектора симметричной квадратной матрицы

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 43.**

1. Интерполяционный многочлен Лагранжа.
2. Исследование устойчивости разностной схемы методом гармоник.
3. Формула погрешности $\delta\lambda$ собственного значения симметричной квадратной матрицы.

Московский Государственный университет им. М.В. Ломоносова
Физический Факультет

ЧИСЛЕННЫЕ МЕТОДЫ

В А Р И А Н Т **№ 44.**

1. Разделенные разности таблично заданной функции.
2. Неявная схема с весами для уравнения колебаний на отрезке. Устойчивость.
3. Получить оценку остаточного члена формулы численного интегрирования трапеций.