Лекция №17. #5.5** Доказательство сходимости разностной схемы к решению задачи Дирихле.

u(x,y)=?

$$u=-f(x,y) \qquad x\in[a,b], \qquad y\in[c,d]$$

$$u(a,y)=\mu_3(y) \qquad u(x,c)=\mu_1(x)$$

$$u(b,y)=\mu_4(y) \qquad u(x,d)=\mu_2(x)$$

Сетка: $x_i=a+ih$, $h=(b-a)/n$; $y_j=b+jk$, $k=(d-a)/m$ (x_i,y_j) – узел сетки

$v=\{v_{ij}\}$, i=0,n; j=0,m -точное решение разностной схемы $u=\{u_{ij}\}$, i=0,n; j=0,m - точное решение (1)

$$\left(v_{x\bar{x}}\right)_{ij}+\left(v_{y\bar{y}}\right)_{ij}=-f_{ij} \qquad i=1,n-1 \quad j=1,m-1$$

$$v_{0j}=\mu_{3j} \quad v_{nj}=\mu_{4j} \quad v_{i0}=\mu_{1i} \quad v_{im}=\mu_{2i}$$

Для отыскания значений функции v в узлах сетки нужно решить линейную систему уравнений относительно вектора $\bar{v}\in R^{(n-1)(m-1)}$, *где* $\bar{v}=\left(v_{11},v_{21},...,v_{n-11},v_{12},v_{22},...,v_{n-12},...,v_{1m-1},v_{2m-1},...,v_{n-1m-1}\right)$

Систему можно записать в матричном виде для сетки 5×5 – (см. распечатку – рис. 2)

$$A=-2\left(\frac{1}{h^2}+\frac{1}{k^2}\right) \qquad A\bar{v}=F \rightarrow \left(Л.15-ф-ма(11)+утв.\right)$$

Было доказано, что detA≠0, $A=A^T$, A<0 и система $-A\bar{v}=-F$ можно решать методами Зейделя и Верхней Релаксации (-A>0).

Докажем теорему о сходимости.

Теорема:

Пусть (решение задачи (1) достаточно гладкое) ⇒ (решение разностной схемы (2) сходиться к решению задачи (1) со 2-м порядком по h и k равномерно, т.е.:

$$(5) \quad \max_{\substack{i=0,n\\ j=0,m}}\left|u_{ij}-v_{ij}\right|\le K\left(h^2+k^2\right), \text{ где } K-\text{не зависит от сетки}$$

Доказательство: 1) изучим свойства матрицы A; 2)получим оценку (5).

#5.5**.1 Принцип Максимума.

Шаблон узла (x_i,y_j): Ш(i,j) – т.е. это набор узлов, которые входят в уравнение, ассоциированное с этим узлом: Ш(3,2)={(2,2),(2±1,2),(2,2±1)}

Ш'(i,j)= Ш(i,j)/(x_i,y_j) – окрестность узла (i,j) Ш'(2,2)= {(2±1,2),(2,2±1)}

Определение:

В теории разностных схем, узел называют граничным, если его окрестность пустая.

Ш(0,1)={(0,1)}, Ш'(0,1)=∅

Среди внутренних узлов выделим узлы:

I-го типа: окрестность которого содержит только внутренние узлы.

II-го типа: окрестность которого содержит хотя бы один граничный узел.

Пример: (x_2,y_2) – внутренний узел 1-го типа.

(x_1,y_1) – внутренний узел 2-го типа, т.к. Ш(1,1)={(1,1),(1±1,1),(1,1±1)} Ш'(1,1)={ (1±1,1),(1,1±1)}

Внутренним узлам в матрице A соответствует строка, в которой 5 ненулевых элементов:

$A,\frac{1}{h^2},\frac{1}{h^2},\frac{1}{k^2},\frac{1}{k^2}$. Внутренним узлам в матрице A соответствует строка в которой у ненулевых элементов u есть коэффициент A.

Новые обозначения: *Пусть* $\bar{v}\in R^{(n-1)(m-1)}$ *– вектор* (3). Запись $\bar{v}\ge 0$ означает, что $v_{ij}\ge 0$

$A\cdot\bar{v}\in R^{(n-1)(m-1)}$ $\left(A\bar{v}\right)_{ij}$ – компонента вектора $A\bar{v}$ с соответствующим индексом.

Запись $A\bar{v}\ge 0$ означает, что $\left(A\bar{v}\right)_{ij}\ge 0,\ i=1,n-1;\ j=1,m-1$

Теорема:

Пусть на сетке (n,m) задана матрица A вида как на рис., то (если для некоторого $\bar{v} \in R^{(n-1)(m-1)}$ $A\bar{v} \geq 0) \Rightarrow \left(\bar{v} \leq 0\right)$

Доказательство: $c = \max\limits_{\substack{i=1,n-1 \\ j=1,m-1}} \left(v_{ij}\right)$. Через (l,s) обозначим индексы компоненты матрицы, на которой достигается максимум: $v_{ls} = \max\limits_{\substack{i=1,n-1 \\ j=1,m-1}} \left(v_{ij}\right) = c$.

Пусть c>0, рассмотрим случай:

1) v_{ls} -внутренний узел 2-го типа
2) v_{ls} -внутренний узел 1-го типа.

Случай 1: пусть $v_{ls}=v_{11}$ и сетка (5,5), тогда из

$$\left(A\bar{v}\right)_{11} = Av_{11} + \frac{1}{h^2}v_{21} + \frac{1}{k^2}v_{12} = \frac{1}{h^2}\underbrace{\left(v_{21} - v_{11}\right)}_{\leq 0} + \frac{1}{k^2}\underbrace{\left(v_{12} - v_{11}\right)}_{\leq 0} - \underbrace{\left(\frac{1}{h^2}v_{11} + \frac{1}{k^2}v_{11}\right)}_{>0} < 0$$

$(Av)_{11}<0$, но по условию теоремы $-A\bar{v} \geq 0 \Rightarrow$ максимум не может достигаться во внутреннем узле 2-го типа.

Случай 2: Пусть v_{ls} – внутренний узел 1-го типа: $\left(Av\right)_{ls} = Av_{ls} + \frac{1}{h^2}v_{l+1s} + \frac{1}{h^2}v_{l-1s} + \frac{1}{k^2}v_{ls+1} + \frac{1}{k^2}v_{ls-1} =$

$$= \frac{1}{h^2}\left(v_{l+1s} - v_{ls}\right) + \frac{1}{h^2}\left(v_{l-1s} - v_{ls}\right) + \frac{1}{k^2}\left(v_{ls+1} - v_{ls}\right) + \frac{1}{k^2}\left(v_{ls-1} - v_{ls}\right) \leq 0$$

$\Rightarrow \left(A\bar{v}\right)_{ls} \leq 0$, но по условию теоремы $\left(A\bar{v}\right)_{ls} \geq 0; \Rightarrow \left(A\bar{v}\right)_{ls} = 0 \Rightarrow$ каждое слагаемое суммы нулевое:

Если положительный максимум достигается во внутреннем узле 1-го типа, то в окрестности этого узла, функция v принимает такое же значение: $\boxed{v_{ls} = v_{l\pm1,s} = v_{ls\pm1}}$.

Случай 1: Пусть в окрестности узла (l,s) есть узел в окрестность которого попадает граничный узел (в окрестности (l,s) есть узел 2-го типа) $\Rightarrow$ получено противоречие.

Пусть в окрестности узла (l,s) все узлы являются узлами 1-го типа. Строим последовательность узлов: $\underbrace{(l,s)}_{вн.уз.1-го\ т.}, \underbrace{(i,j)_1, (i,j)_2, \ldots, (i,j)_t}_{не\ важно\ какие}$ $\underbrace{}_{вн.уз.2-го\ т.}$.

Важно, что $(i,j)_1 \in Ш'(l,s), (i,j)_2 \in Ш'\left((i,j)_1\right), \ldots, (i,j)_t \in Ш'\left((i,j)_{t-1}\right)$.

Значит $v_{ls} = c, v_{(i,j)1} = c, v_{(i,j)2} = c, \ldots, \boxed{v_{(i,j)t} = c}$ - но этого быть не может.

Следствие 1:

Матрица разностной схемы A такова, что $\left(A\bar{v} \geq 0\right) \Rightarrow \left(\bar{v} \leq 0\right), \left(A\bar{v} \leq 0\right) \Rightarrow \left(\bar{v} \geq 0\right)$

Следствие 2: detA≠0

Доказательство: рассмотрим $A\bar{v} = 0$ (*) $\left\{\left(A\bar{v}\right) \geq 0 \Rightarrow \bar{v} \leq 0; \left(A\bar{v}\right) \leq 0 \Rightarrow \bar{v} \geq 0\right\}$ $\bar{v} = 0$

Следствие 3:

Рассмотрим $A\bar{v} = F$. Если F≥0$\Rightarrow \bar{v}$ - решение системы и $\bar{v}$≤0.

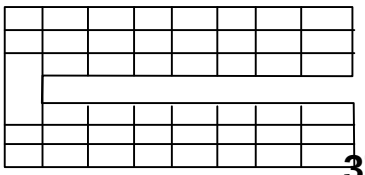
Определение:

Сетка называется связной, если для любой пары узлов, один из которых имеет не пустую окрестность можно построить последовательность узлов, соединяющих эти узлы через окрестность в следующем смысле:

(l,s) – начальный узел Ш'(l,s)≠∅ (l',s') – конечный узел

$(i,j)_1, (i,j)_2, \ldots, (i,j)_t$ – последовательность узлов для связки

$(i,j)_1 \in Ш'(l,s), (i,j)_2 \in Ш'((i,j)_2), \ldots$ $(l',s') \in Ш'((i,j)_t)$



пример не связной сетки

при доказательстве принципа максимума мы использовали связность сетки.

Лекция №18.

Доказательство сходимости решения разностной схемы к решению задачи Дирихле.

Рассмотрим задачу (1) на равномерной сетке в прямоугольнике и разностную схему (2).

$u=\{u_{ij}\}$, i=0,n; j=0,m - точное решение дифференциальной задачи (1) в узлах сетки.

$v=\{v_{ij}\}$, i=0,n; j=0,m - точное решение разностной схемы (2) в узлах сетки.

$z=\{z_{ij}\}$, i=0,n; j=0,m - погрешность решения дифференциальной задачи с помощью разностной схемы. $z_{ij} \overset{def}{=} u_{ij} - v_{ij}$

$\psi=\{\psi_{ij}\}$, i=0,n; j=0,m - погрешность аппроксимации задачи (1) схемой (2).

$\psi_{ij} \overset{def}{=}$ это невязка уравнения разностной схемы с номером i,j, при условии, что туда подставили точное решение задачи (1).

$$\psi_{ij} \underset{\substack{i=1,n-1\\ j=1,m-1}}{=}$$

$$\psi_{0j} = u_{0j} - \mu_{1j} = 0 \quad \psi_{nj} = u_{nj} - \mu_{2j} = 0 \quad \psi_{i0} = 0 \quad \psi_{im} = 0$$

Раньше была получена оценка:

$$(6)\max_{\substack{i=0,n\\ j=0,m}}\left|\psi_{ij}\right| \le \underbrace{\frac{1}{12}\max_{\substack{x\in[a,b]\\ y\in[c,d]}}\left(\left|\frac{\partial^4 u}{\partial x^4}\right|,\left|\frac{\partial^4 u}{\partial y^4}\right|\right)}_{=M}\left(h^2+k^2\right) \Rightarrow \max_{\substack{i=0,n\\ j=0,m}}\left|\psi_{ij}\right| \le M\left(h^2+k^2\right)$$

Воспользуемся связью величин z и ψ:

$$\begin{cases} \left(z_{\bar{x}x}\right)_{ij} + \left(z_{\bar{y}y}\right)_{ij} = \psi_{ij} \quad (i=1,n-1;\ j=1,m-1) \\ гран.усл.:\ z_{0j} = \psi_{0j} = 0,\ z_{nj} = \psi_{nj} = 0 \\ z_{i0} = \psi_{i0} = 0, \quad z_{im} = \psi_{im} = 0 \end{cases} \qquad (7)$$

Идея доказательства сходимости: построить уравнение такого же типа, как уравнение (7), правая часть которого будет мажорировать правую часть (7). Затем, используя принцип максимума покажем, что решение нового уравнения будет мажорировать решение уравнения (7).

Оценим $|z_{ij}|\le$? и докажем сходимость. Оценим z через ψ и воспользуемся оценкой для ψ.

$\left|z_{ij}\right| \le M\left|\psi_{ij}\right| \le M M\left(h^2+k^2\right) \underset{h,k\to 0}{\to} 0$ Будем строить мажорирующее уравнение:

рассмотрим $\overset{B}{z}(x,y) \overset{def}{=} \frac{K}{4}\left((x-a)(b-x)+(y-c)(d-y)\right)$ Определим $K \overset{def}{=} M\left(h^2+k^2\right)$, где M из (6).

Свойство 1: при x∈[a,b]; y∈[c,d]⇒$\overset{B}{z}(x,y) \ge 0$

Свойство 2: $K \ge |\psi_{ij}|$, i=0,n; j=0,m.

Свойство 3: построим задачу Дирихле такую, чтобы $\overset{B}{z}$ было ее решением (в прямоугольнике [a,b]×[a,b].

$$(8)\begin{cases} \overset{B}{z} = -K \le 0 \\ \overset{B}{z}(a,y) = \overset{B}{z}(b,y) = \frac{K}{4}(y-c)(d-y) \ge 0 \\ \overset{B}{z}(x,c) = \overset{B}{z}(x,d) = \frac{K}{4}(x-a)(b-x) \ge 0 \end{cases}$$

Свойство 4: на равномерной сетке прямоугольника [a,b]×[a,b] функция $\overset{B}{z}$ является решением следующей разностной схемы:

$$(9)\begin{cases}\left(\hat{z}_{x\bar{x}}\right)_{ij}+\left(\hat{z}_{y\bar{y}}\right)_{ij}=-K\le 0 \quad i=1,n-1; \quad j=1,m-1\\ \hat{z}_{0j}=\hat{z}_{nj}=\dfrac{K}{4}\left(y_j-c\right)\left(d-y_j\right)\ge 0 \qquad j=1,m-1\\ \hat{z}_{i0}=\hat{z}_{im}=\dfrac{K}{4}\left(x_i-a\right)\left(d-x_i\right)\ge 0 \qquad i=1,n-1\end{cases}$$

<u>Доказательство:</u> $\hat{z}$(x,y) – есть полином 2-го порядка от x и y ⇒ решение разностной схемы и задачи Дирихле совпадают.

<u>Замечание:</u> модули правых частей системы (9) больше модулей правых частей системы (7), т.к. $|\psi_{ij}|\le K$. Чтобы использовать это свойство перепишем системы (7) и (9) в матричном виде и используем принцип Максимума. Запишем (7) и (9) как системы линейных уравнений относительно вектора размерности (n-1)(m-1) только для внутренних узлов.

(7*) Система для $\bar{z}$ имеет вид: $A\bar{z}=F_1$. (9*) Система для $\bar{\hat{z}}$ имеет вид: $A\bar{\hat{z}}=F_2$. $\bar{z}\in R^{(n-1)(m-1)}$, $\bar{\hat{z}}\in R^{(n-1)(m-1)}$ – на внутренних узлах.

Для краткости вместо $\bar{\hat{z}}$ будем писать $\hat{z}$:

$$(10)\ A\left(\hat{z}+z\right)=F_1+F_2 \qquad (11)\ A\left(\hat{z}-z\right)=F_1-F_2$$

Запишем систему (7*) для сетки 5×5:

$$(7*)\begin{bmatrix} & & \\ & A & \\ & & \end{bmatrix}\times\begin{pmatrix}z_{11}\\ z_{21}\\ \vdots\\ z_{44}\end{pmatrix}=\begin{pmatrix}\psi_{11}\\ \psi_{21}\\ \vdots\\ \psi_{44}\end{pmatrix}$$

Каждая компонента правой части этой системы состоит из одного слагаемого

Система (9*) в матричном виде для сетки 5×5:

$$(9*)\begin{bmatrix} & & \\ & A & \\ & & \end{bmatrix}\times\begin{pmatrix}\hat{z}_{11}\\ \hat{z}_{21}\\ \vdots\\ \hat{z}_{44}\end{pmatrix}=\begin{pmatrix}-K-\frac{1}{h^2}\cdot\hat{z}_{10}-\frac{1}{k^2}\hat{z}_{01}\\ -K+\ldots\\ \vdots\\ -K+\ldots\end{pmatrix}$$

некоторые из компонент правой части будут иметь по два или по три слагаемых

⇒ каждая компонента системы (9*) состоит из отрицательных (неположительных) слагаемых.

Перейдем к анализу (10), (11):

Рассмотрим $A\left(\hat{z}\pm z\right)$ как вектор размерности $R^{(n-1)(m-1)}$ $\left(A\left(\hat{z}\pm z\right)\right)_{ls}$ - его компоненты с индексами

- Если $\left(A\left(\hat{z}\pm z\right)\right)_{ls}=\pm\psi_{ls}-K\le 0$ – по свойству 2.
- Если узел (l,s) – является внутренним узлом 2-го типа, то

$$\left(A\left(\hat{z}\pm z\right)\right)_{ls}=\pm\psi_{ls}\underbrace{-K}_{\le 0}-\underbrace{\ldots}_{слаг.\ из\ гран.\ усл.}\le 0$$

<u>Итог:</u>

$$\boxed{\forall\ узла\ (l,s):\ A\left(\hat{z}\pm z\right)\le 0}\quad (12)$$

По принципу максимума $\hat{z}\pm z\ge 0$. Аналогично исследуем систему (11) для вектора $\left(\hat{z}-z\right)$.

$$\Rightarrow \begin{cases} \hat{z} + z \ge 0 & z \ge -\hat{z} \\ \hat{z} - z \ge 0 & z \le \hat{z} \end{cases} \text{ и } \hat{z} \ge 0 - \text{по св-ву 1} \Rightarrow -\hat{z} \le z \le \hat{z} \Rightarrow |z| \le \hat{z} \, (13)$$, то есть каждая компонента вектора z по модулю не превосходит соответствующей компоненты вектора $\hat{z}$, соответствующих внутренним узлам: $|z_{ij}| \le \hat{z}_{ij}$, $i = 1, n-1$; $j = 1, m-1$; $\max\limits_{\substack{i=0,n \\ j=0,m}} |z_{ij}| \le \max\limits_{\substack{i=1,n-1 \\ j=1,m-1}} |z_{ij}| \le \max\limits_{\substack{i=1,n-1 \\ j=1,m-1}} |\hat{z}_{ij}| \underset{т.к.\ \hat{z} \ge 0}{=}$

$$= \max_{\substack{i=1,n-1 \\ j=1,m-1}} z_{ij} \le \frac{K}{4}\left(\frac{(b-a)^2}{4} + \frac{(d-c)^2}{4}\right) = \frac{M\left((b-a)^2 + (d-c)^2\right)}{16}\left(h^2 + k^2\right) \Rightarrow$$

$$\Rightarrow \max_{\substack{i=0,n \\ j=0,m}} |z_{ij}| = \underbrace{\frac{M}{16}\left((b-a)^2 + (d-c)^2\right)}_{=K}\left(h^2 + k^2\right) \, (14)$$

$\Rightarrow$ мы доказали сходимость решения разностной схемы к решению задачи Дирихле со 2 порядком. Формула (14) – это и есть формула (5) которую мы хотели доказать (Л.15).

1. Замечание: при получении оценки (14) мы использовали следующее свойства:

свойство: f(x)=(x-a)(b-x). Очевидно, что $\max\limits_{x \in [a,b]} |f(x)| = f\left(\frac{a+b}{2}\right) = \left(\frac{b-a}{2}\right)^2 = \frac{(b-a)^2}{4}$

2. Замечание: оценка позволяет оценить константу M для тестовой задачи: позволяет выбрать сетку, для того, чтобы погрешность z была достаточно малой. Оценка (14) полезна в 2-х аспектах:

1) доказывает сходимость.
2) во время отладки программы или метода для решения тестовой задачи, можно оценить какой должна быть сетка, чтобы точное решение разностной схемы отличалось от точного решения дифференциальной задачи не более, чем на некоторый ε, заранее заданный. (для тестовой задачи можно оценить величину M.)

3. Замечание: оценку (14) можно сделать более точной, если для $\max\limits_{\substack{i=0,n \\ j=0,m}} |\psi_{ij}|$ вместо оценки (6)

использовать оценку: $\max\limits_{\substack{i=0,n \\ j=0,m}} |\psi_{ij}| \le \frac{1}{12}\left(\max\limits_{\substack{x \in [a,b] \\ y \in [a,b]}} \left|\frac{\partial^4 u}{\partial x^4}\right| h^2 + \max\limits_{\substack{x \in [a,b] \\ y \in [a,b]}} \left|\frac{\partial^4 u}{\partial y^4}\right| k^2\right)$

Лекция №19. #5.6 Итерационные методы Линейной Алгебры.

1. Постановка задачи.
2. Описание методов.
3. Общие свойства методов.
4. Особенности применения метода для задачи Дирихле.

1. Постановка задачи.

(1) Ax=b, $x \in R^{(n)}$, $A=A^T>0$; $\lambda_i(A)$, i=1,n – собственные числа A.

Если матрица положительно определена и симметрична, то все ее собственные числа $\lambda_i>0$, i=1,n

$0 < \lambda_1 \le \lambda_2 \le \ldots \le \lambda_{n-1} \le \lambda_n$ – максимальное; $\lambda_1 = \min\limits_{i=1,n} \lambda_i(A)$; $\lambda_n = \max\limits_{i=1,n} \lambda_i(A)$

Определение: $\mu = \mu(A) \overset{def}{=} \frac{\lambda_n}{\lambda_1}$ – число обусловленности матрицы A, согласованное с евклидовой нормой.

x* - точное решение системы (1).

Вид итерационной системы: $B_s \frac{x^{(s+1)} - x^{(s)}}{\tau_s} + Ax^{(s)} = b$; $x^{(0)}$ – начальное приближение;

Параметры метода: B_s – матрица; τ_s – число.

Пусть $x^{(s)} \sim x_s$ – эквивалентные обозначения.

Определение критерия остановки и сходимости метода смотрите в декабре.

Примеры одношаговых методов:

Зейдель, Якоби, Верхней Релаксации.

2. Метод простой итерации (МПИ).

Общий вид метода: $\frac{x_{s+1}-x_s}{\tau}+Ax=b$ (3)

Теорема:

Если $A=A^T>0$ и $\tau\in(0,2/\lambda_n)$, то метод (3) сходиться к x^* ($\forall x_0$)

Метод с оптимальным параметром: $\tau_{опт}=\frac{2}{\lambda_n+\lambda_1}$ (4)

Теорема:

Если $A=A^T>0$, то в классе методов (3) наиболее быструю сходимость имеет метод с параметром (4)

Справедлива оценка: $\|x_s-x^*\|\le(\varphi(\mu_A))^s\|x_0-x^*\|$, где $\varphi(\mu_A)$ – число зависящее от свойств матрицы: $\varphi(\mu_A)\in(0,1)$.

Следствие: при $s\to\infty$ $\underbrace{\|x_s-x^*\|}_{\to 0}\le\underbrace{(\varphi(\mu_A))^s}_{\to 0,\,s\to\infty}\|x_0-x^*\|$ (5). Эта оценка полезна в теоретическом плане, т.к. из нее следует сходимость; x^* - неизвестно, поэтому (5) бесполезна в практическом смысле.

3. Метод простой итерации с чебышевским набором параметром.

Такой метод имеет $K>1$ параметров: $\tau_0, \tau_1,\ldots,\tau_{K-1}$ (K – задается пользователем)

Все τ_s можно вычислить заранее, до расчетов по методу. Непостоянные τ_s вводятся для повышения скорости сходимости.

$$\frac{x_{s+1}-x_s}{\tau_s}+Ax_s=b\ (6)\quad \tau_s=\frac{1}{\frac{\lambda_1+\lambda_n}{2}+\frac{\lambda_n-\lambda_1}{2}\cos\left(\frac{\pi}{2K}(1+2s)\right)},\ s=0,K-1$$

x_0 – начальное приближение. τ_0 используется для вычисления x_1, $\tau_1\to x_2,\ldots,\tau_{K-1}\to x_K$. Далее продолжаем с той же последовательностью: $x_{0new}=x_K$; исп. τ_0 для $x_{1new}=x_{K+1}$; исп. τ_1 для $x_{2new}=x_{K+2}$ и т.д. Таким образом МПИ-Чеб используется для расчета приближений порциями по К.

x_K лучше x_0, x_{2K} лучше $x_K,\ldots$

Внутри же каждого цикла с расчетом $x_1,\ldots,x_K$; $x_{K+1},\ldots,x_{2K}$; и т.д. метод может давать не мотонную погрешность, но последовательность $x_K, x_{2K}, x_{3K},\ldots,x_{jK}\to x^*, j\to\infty$

Теорема:

МПИ-Чеб. является оптимальным в следующем смысле: пусть $A=A^T>0$; дано целое $K>1$. Тогда $\forall x\in R^n, \forall A=A^T$, таких, что спектр матрицы A находиться в тех же границах, что и спектр матрицы A: $\lambda_i(A)\in[\lambda_1(A),\lambda_n(A)]$, метод дает наилучшую гарантию убывания погрешности через К итераций: $\|x_K-x^*\|\le 2(\varphi_{МПИ-Чеб}(\mu_A))^K\cdot\|x_0-x^*\|$ (7), *где* $\varphi(\mu_A)\in(0,1)$

Из оценки (7) следует сходимость: $\underbrace{\|x_{jK}-x^*\|}_{\to 0,\,j\to\infty}\le 2\underbrace{(\varphi(\mu_A))^{jK}}_{\to 0}\cdot\|x_0-x^*\|$

Отдельно взятые методы класса (6), для отдельно взятых x_0 и отдельно взятых матриц A с $\lambda_i(A)\in[\lambda_1,\lambda_n]$ могут дать лучший результат и за меньшее число итераций, чем Чеб. метод, просто Чеб. метода дает наилучшую гарантию в некотором среднем случае (при отсутствии другой априорной информации).

4. Метод минимальных невязок.

$\frac{x_{s+1}-x_s}{\tau_s}+Ax_s=b$, значение τ_s заранее неизвестно, оно будет известно, когда буде найдено x_s:

$\tau_s=\frac{(Ar_s,r_s)}{(Ar_s,Ar_s)}$,(10) $r_s \overset{def}{=} Ax_s-b$ – невязка системы (1) при подстановке x_s

Теорема:

Если $A=A^T>0$, то $\|x_s-x^*\|\le\mu_A(\varphi_{мн}(\mu_A))^s\|x_0-x^*\|$ (9), где $\varphi(\mu_A)\in(0,1)$

Из (9) следует сходимость метода минимальных невязок. Чтобы использовать метод ненужно знать собственные числа матрицы А.

5. Метод сопряженных градиентов.

Вместо системы (1) решается оптимизационная задача: $F(x)=(Ax,x)-z(b,x)\to\min\limits_{x\in R^n}$

x_0 - начальное приближение x_s – приближение шага s; $r_0=Ax_0-b$; $r_s=Ax_s-b$

Метод: $\forall x_0\in R^n \quad x_1=x_0+\alpha_0h_0,\ h_0\in R^n-вектор,\ \alpha_0-число,\ h_0=-r_0=-(Ax_0-b)$

$$\alpha_0=\frac{(r_0,h_0)}{(Ah_0,h_0)}=\frac{(Ax_0-b,h_0)}{(Ah_0,h_0)};\quad x_{s+1}=x_s+\alpha_sh_s,\ \alpha_s-число,\ h_s\in R^n-вектор$$

$$h_s=-r_s+\beta_sh_{s-1},\ где\ r_s=Ax_s-b,\ h_{s-1}-с\ предыдущего\ шага$$

$$\beta_s=\frac{(Ah_{s-1},r_s)}{(Ah_{s-1},h_{s-1})}\quad \alpha_s=\frac{(r_s,h_s)}{(Ah_s,h_s)}=\frac{(Ax_s-b,h_s)}{(Ah_s,h_s)}$$

x_{s+1} зависит от невязки шага r_{s+1} и h_{s-1}, которые используются при вычислении x_s (отсюда этот метод двушаговый).

Свойства метода.

1) Для реализации метода не нужно знать собственные числа матрицы А.
2) Через определенное число шагов метод даст более хороший результат, чем метод Гаусса.

Теорема:

Если $A=A^T>0$, то $\forall x_0\in R^n$, либо x_0 – точное решение x^*, либо $\exists K\in\{1,2,\ldots,n\}$, что $x_K=x^*$ - есть точное решение. (Метод сопряженных градиентов – прямой)

3. Теорема:

Если $A=A^T>0$, то $\forall x_0\in R^n$ и $\forall s=0,1,\ldots,n-1$ выполняется оценка: $\|x_s-x^*\|_A\le(\varphi_{сг}(\mu_A))^s\|x_0-x^*\|_A$,

где $\|A\|_A\overset{def}{=}\sqrt{(Ax,x)}$ – энергетическая норма, порожденная матрицей А, при этом $x_n=x^*$ - точ.реш.

Критерии остановки и продолжения счета по методу сопряженных градиентов.

теоретически, если $x_s=x^*$, то $r_s=Ax_s-b=0\Rightarrow\beta_s=0$, $\to h_s=-r_s+\beta_sh_{s-1}=0$; $x_{s+1}=x_s+\alpha_sh_s=x_s$.

Для α_s нежен $h_s=0$, $\alpha_s=\ldots/(Ah_s,h_s)=1/0$ (ошибка).

Теоретически здесь происходит остановка, поэтому на каждом шаге метода проверяется условие, чтобы невязка $r_s\ne0$. Если $r_s=0\Rightarrow x_s=x^*$ - остановка.

Если $x_s\approx x^*$, то $r_s\approx0$ и h_s, $\beta_s\approx0$ и вычислительные погрешности счета мешают погрешности метода.

Если r_s достаточно мала, то x_s принимается за начальное приближение: $x_{s+1}=x_{1new}+\alpha_0h_{0new}$.

Лекция №20. #6 Сравнение методов МПИ, МПИ-Чеб, МН, СГ.

1) Методы применяются для решения матрицы $A=A^T>0$.
2) В методах МПИ и МПИ-Чеб нужно знать собственные числа матрицы А.
3) Для методов МН и СГ собственные числа не нужны, но параметры метода приходиться вычислят в процессе счета.
4) Если пренебречь деталями: $\|x_s-x\|_{мет}\le\alpha_{мет}\cdot(\varphi_{мет}(\mu_A))^s\|x_0-x^*\|_{мет}$
5) Если $\mu_A\to\infty$, то $\varphi(\mu_A)\to1$. Если $\mu_A>>1$, все методы сходятся медленно.

6) Величины φ и α будут выведены в #7.
7) МПИчеб при больших К может оказаться вычислительно – неустойчивым, поэтому в реализации метода используют специальный порядок чередования τ_s (Бахвалов).

Лабораторная работа №3.

Разностную схему задачи Дирихле мы записывали в виде: $A\bar{v} = F \quad (*)$, $\bar{v}$ - значение точного решения разностной схемы во внутренних узлах сетки.

Утверждение: $A=A^T<0$.

Вывод: систему $-A\bar{v} = -F \quad (**)$ можно решить с помощью МПИ, МПИчеб, СГ, МН, т.к. $(-A)=(-A)^T>0$.

Утверждение: λ_{ij} – собственные числа матрицы А:

$$\lambda_{ij}(A) = \frac{4}{h^2}\sin^2\left(\frac{\pi i}{2h}\right) + \frac{4}{k^2}\sin^2\left(\frac{\pi j}{2m}\right), \quad i = 1, n-1;\ j = 1, m-1$$

Вывод: $\lambda_{\min} = \min\left(\lambda_{ij}(-A)\right) = \lambda_{11}$; $\lambda_{\max} = \max\left(\lambda_{ij}(-A)\right) = \lambda_{n-1m-1}$; $\mu_A = \dfrac{\lambda_{n-1m-1}}{\lambda_{11}}$

Утверждение: (n,m) – размерность сетки, тогда при $n\to\infty$, $m\to\infty$, $\mu(-A)\to\infty$.

Пример: Задача Дирихле на квадрате: $x\in[0,1]$, $y\in[0,1]$. Сетка n=m. h=k=1/n.

$$\lambda_{\min} = \frac{8}{h^2}\sin^2\frac{\pi}{2n}; \quad \lambda_{\max} = \frac{8}{h^2}\sin^2\frac{\pi(n-1)}{2n} = \frac{8}{h^2}\cos^2\frac{\pi}{2n}; \quad \mu(-A) = ctg^2\left(\frac{\pi}{2n}\right) \sim \left(\frac{2n}{\pi}\right)^2 \approx \frac{1}{2}n^2,\ n\to\infty$$

сетка	(10,10)	(100,100)	(1000,1000)
$\mu(-A)$	50	$0.5\cdot10^4$	$0.5\cdot10^6$

Метод Верхней Релаксации.

A=L+D+R

$$(D+\omega L)\frac{x_{s+1}-x_s}{\omega} + Ax_s = b, \quad x_0 \in R^n$$

оценка оптимального ω имеет вид: $\omega_{opt} \approx \dfrac{2}{1+\sqrt{1-\rho^2(B)}}$,

$B = D^{-1}\cdot(L+R)$, $\rho(B) = \max\limits_{i=1,n}\left\{|\lambda_i(B)|\right\}$ - спектральный радиус В.

Принцип максимума для разностных схем.

1) Запись РС в канонической форме.
2) Теорема и условия принципа максимума (Пmax).
3) Примеры.

1. Пусть $x\in R^N$ и на области $G\subset R^N$ дана некоторая линейная дифференциальная задача с граничными условиями: $Lu(x)=f(x)$ при $x\in G$. $lu(x)|_{\partial G}=\mu(x)$ $u(x)=?$

Ω_H – сетка на $\bar{G}$, H – параметры сетки (h,k), z – узел сетки.

Полагаем, что каждому узлу сетки ассоциировано только одно уравнение.

Ш(x) – это набор узлов, участвующих в уравнении, ассоциированном с узлом x.

Ш'(x) – окрестность Ш(x): Ш'(x)= Ш(x)/{x}

$$(1)\begin{cases} A(x)v(x) - \sum\limits_{\xi\in Ш'(x)} B(x,\xi)v(\xi) = \Phi(x), x\in\Omega_H \\ \xi\in Ш'(x) \end{cases}$$

– это каноническая формула записи разностных схем

v(x) – сеточная функция, являющаяся решением разностной схемы.

A(x) – коэффициент, с которым узел x участвует в своем уравнении.

B(x,ξ) – коэффициент, с которым узел ξ, участвует в уравнении, ассоциированном с узлом x.

Φ(x) – правая часть уравнения, ассоциированного с узлом x.

Введем линейный оператор L: действующий на сеточную функцию v(x):

$$(2)\ \boxed{L(x)v(x) \overset{def}{=} A(x)v(x) - \sum_{\xi \in Ш'(x)} B(x,\xi)v(\xi)}$$

Таким образом, $\boxed{L(x)v(x) = \Phi(x), x \in \Omega_H}$ (3) - это операторная форма записи разностной схемы.

2. В доказательстве теоремы (п.5) мы пользовались тем, что сетка связная; строка матрицы A, соответствующая внутреннему узлу 1-го типа имеет свойство нестрогого диагонального преобладания. И строка матрицы A, соответствующая узлу 2-го типа имеет свойство строго диагонального преобладания (пр: стр. 1 и 7).
Сформируем более точный, абстрактный аналог этих свойств.

Определение:

Сетка Ω_H называется связной, если для любых узлов x,y можно построить последовательность x_1, x_2, …, x_t, такую, что $x_1 \in Ш'(x)$, $x_2 \in Ш'(x_1)$, …,$y \in Ш'(x_t)$.

Определение:

Оператор L(x) удовлетворяет условию сильной положительности (++) на Ω_H, если:

$\forall x \in \Omega_H \quad A(x) > 0 \ (4) \qquad \forall \xi \in Ш'(x) \quad B(x,\xi) > 0 \ (5)$

$A(x) - \sum_{\xi \in Ш'(x)} B(x,\xi) \geq 0 \ (6)$ *и если* $\exists x^* \in \Omega_H$, *что* $A(x^*) - \sum_{\xi \in Ш'(x^*)} B(x^*,\xi) > 0 \ (7)$

Если для узла x: Ш'(x)=∅, то (5) не проверяется, а (6) и (7) следует читать так: $A(x) \geq 0$ *и* $A(x^*) > 0$, таким образом, для узла с пустой окрестностью нужно проверить только условие (4).

Теорема: Принцип Максимума.

Если Ω_H – связная и L(x) удовлетворяет условию сильной положительности, то из (L(x)v(x)≥0) ⇒ ⇒(v(x)≥0 для $\forall x \in \Omega_H$).

Следствие 1:

(Если Ω_H – связная и L(x) ++) ⇒(то при ∀Φ(x) разностная схема (3) имеет единственное решение. (detA≠0 #5: Av≥0, v≤0)

Следствие 2:

Пусть на связной сетке Ω_H заданы: $(8)\ L(x)v(x) = \Phi(x) \quad (9)\ L(x)\hat{v}(x) = \Phi(x)$, такие, что: L(x) ++ и $\forall x \in \Omega_H \ |\Phi(x)| \leq \Phi(x)$, Тогда, $\forall x \in \Omega_H \ |v(x)| \leq \hat{v}(x)$ (пользовались при док-ве схемы в #5)

Иногда удобно, уравнение (9) называть мажорирующим, т.к. $\Phi(x) > \Phi(x)$ и его решение > решения (8).

Следствие 3:

Если разностная схема (3) заданы на связной сетке Ω_H, L(x) ++ ⇒ схема однозначно разрешима ∀Φ(x).

И для доказательства сходимости разностной схемы (3) к решению дифференциальной задачи (*) можно использовать следующий прием:

1) Оценить ψ(x) (погрешность аппроксимации).
2) Составить уравнение L(x)z(x)=ψ(x).
3) Подобрать мажорирующее уравнение:

$$L(x)\hat{z}(x) = \psi(x), \text{где } \psi(x) \geq |\psi(x)|, \text{тогда } |z(x)| \leq \hat{z}(x), \ \forall x \in \Omega_H \ (10)$$

Мажорирующее уравнение подбирается так, чтобы $\hat{z}$ - было известно.

Практические замечания:

Основное уравнение дифференциальной задачи (*) и граничное условие задачи (*), как правило на сетках аппроксимируются разными выражениями.

Как правило, на сетках в соответствии с физическим смыслом задачи различают внутренние и граничные узлы – различают их по топологическому принципу.

Определение:

Узел называется топологически внутренним, если Ш'(x)≠∅. Узел называется топологически граничным, если Ш'(x)=∅.

ω_H - внутренние узлы сетки; γ_H – граничные узлы сетки; $\Omega_H=\omega_H \cup \gamma_H$ (11).

$$\begin{cases} A((x)v(x) - \sum\limits_{\xi \in Ш'(x)} B(x,\xi)v(\xi) = \Phi(x), x \in \omega_H \quad (1^*) \\ A((x)v(x) = \Phi(x),\ x \in \gamma_H \end{cases}$$

Проверка ++ для L(x):

1) Если Ω_H такая, что $\gamma_H=0$, то $\forall x \in \omega_H$ проверяем (4), (5), (6) и ищем $x^* \in \omega_H$, чтобы выполнялось (7).
2) Если Ω_H таков, что $\gamma_H \neq 0$, то $\forall x \in \omega_H$ проверяем (4),(5),(6) и для $\forall x \in \gamma_H$ проверяем условие (4).

(Д/з обосновать эту схему)

3. Пример: $\begin{cases} u''(x) = \cos(x), \quad x \in (0,1) \\ u(0) = 1 \ \ u \ \ u(1) = 9 \end{cases}$ $\quad \Omega_H = \left\{ x_i = ih,\ i = 0,n;\ h = \dfrac{1}{n} \right\}$

$$\begin{cases} \dfrac{v_{i-1} - 2v_i + v_{i+1}}{h^2} = \cos x_i \\ v_0 = 1 \quad v_n = 9 \end{cases}$$

$Ш_i=\{i,i\pm1\}, i=1,n-1$ $\quad Ш_0=\{0\}$, $\quad Ш_n=\{n\}$. $\quad Ш'_i=\{i\pm1\}$, $Ш'_0=0$, $Ш'_n=0$

$\omega_H=\{x_i, i=1,n-1\}$ $\quad \gamma_H=\{x_0,x_n\}$

Проверим ++:

$$\begin{cases} \dfrac{2}{h^2}v_i - \dfrac{1}{h^2}v_{i+1} - \dfrac{1}{h^2}v_{i-1} = -\cos x_i, \quad i = 1, n-1 \\ 1 \cdot v_0 = 1, \quad 1 \cdot v_n = 9 \end{cases}$$

$$A(i) = \frac{2}{h^2},\ B(i,i+1) = \frac{1}{h^2},\ B(i,i-1) = \frac{1}{h^2},\ \Phi(i) = \cos x_i,\ A(0) = 1,\ \Phi(0) = 1,\ A(n) = 1,\ \Phi(n) = 9$$

Проверим для $\forall x \in \omega_H$:

$$A > 0, B > 0,\ A - \sum B = \frac{2}{h^2} - \frac{1}{h^2} - \frac{1}{h^2} = 0 \geq 0$$

для $x \in \gamma_H$: $A > 0 \Rightarrow L(x)$ – удовлетворяет ++ и РС (13) удовлетворяет Пmax

Пример: $\begin{cases} u''(x) - 3u(x) = 7, \quad x \in (0,1) \\ u'(0) = 3 \quad u'(1) = 9 \end{cases}$ $\quad \Omega_H = \left\{ x_i = ih,\ i = 0,n;\ h = \dfrac{1}{n} \right\}$

$$\begin{cases} \dfrac{v_{i-1} - 2v_i + v_{i+1}}{h^2} - 3v_i = 7 \\ \dfrac{v_1 - v_0}{h} = 3 \quad \dfrac{v_n - v_{n-1}}{h} = 9 \end{cases}$$

$Ш_i=\{i,i\pm1\}, i=1,n-1$ $\quad Ш_0=\{0,1\}$, $\quad Ш_n=\{n,n-1\}$. Граничных узлов нет!!!

$\omega_H=\{x_i, i=1,n\}=\Omega_H$

$$\begin{cases} \overset{A(i)}{\left(\frac{2}{h^2}+3\right)} v_i - \overset{B(i,i+1)}{\frac{1}{h^2}} v_{i+1} - \overset{B(i,i-1)}{\frac{1}{h^2}} v_{i-1} = \overset{\Phi(i)}{-7} \\ \underset{A(0)}{\frac{1}{h}} v_0 - \underset{B(0,1)}{\frac{1}{h}} v_1 = \underset{\Phi(0)}{-3} \qquad \underset{A(n)}{\frac{1}{h}} v_n - \underset{B(n,n-1)}{\frac{1}{h}} v_{n-1} = \underset{\Phi(n)}{9} \end{cases}$$

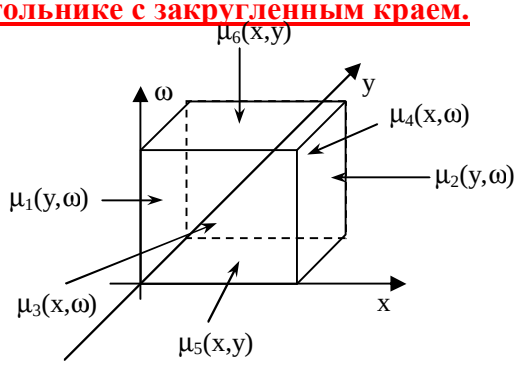
Все выполняется!!!

Лекция №21.

Задачи Дирихле в 3D-области. Дирихле на прямоугольнике с вырезанным квадратами, Дирихле на прямоугольнике с закругленным краем.

1. Задача Дирихле в кубе.

$$(16)\begin{cases} u(x,y.\omega) = -f(x,y,\omega) \\ (x,y,\omega) \in (0,1)\times(0,1)\times(0,1) = G \\ u(x,y,z)\big|_{\partial G} = \mu(x,y,\omega) \end{cases}$$



$\mu(x,y,\omega)=\{\mu_1,\mu_2,\ldots,\mu_6\}$

Сетка: (n,m,p)

$x_i=ih$, $h=1/n$

$y_j=jk$, $k=1/m$

$\omega_l=ls$, $l=1/p$

$u(x,y,\omega)=\{u_{ijl}\}$, i=0.n; j=0,m; s=0,p – сеточная функция – значение точного решения задачи (16) в узлах сетки.

$v(x,y,\omega)=\{v_{ijl}\}$, i=0.n; j=0,m; s=0,p – точное решение разностной схемы в узлах сетки

$z_{ijl}=u_{ijl}-v_{ijl}$- погрешность решения дифференциальной задачи с помощью разностной схемы.

Построим разностную схему 2-го порядка аппроксимации, заменяя 2-е частные производные 3-х точечными разностным оператором 2-й разностной производной.

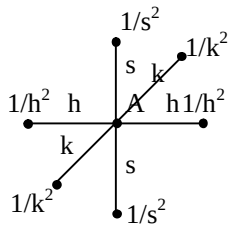
$$(17)\begin{cases} \left(v_{x\bar{x}}\right)_{ijl} + \left(v_{y\bar{y}}\right)_{ijl} + \left(v_{\omega\bar{\omega}}\right)_{ijl} = -f_{ijl} & i=1,n-1;\ j=1,m-1;\ l=1,p-1 \quad (17*) \\ v_{ij0} = \mu_{5ij};\quad v_{ijp} = \mu_{6ij};\quad v_{i0l} = \mu_{3il}; & i=0,n;\ j=0,m;\ l=0,p \\ v_{iml} = \mu_{4il};\quad v_{0jl} = \mu_{1jl};\quad v_{njl} = \mu_{2jl}; & i=0,n;\ j=0,m;\ l=0,p \end{cases}$$

Шаги: {h по x; k по y;s по ω} . Размеры сетки (n,m,p). Индексы i,j,l.

Из физического смысла задачи предполагается, что μ согласуются между собой и на ребрах куба дадут одинаковый результат.

(17*) это: $\dfrac{v_{i-1jl}-2v_{ijl}+v_{i+1jl}}{h^2} + \dfrac{v_{ij-1l}-2v_{ijl}+v_{ij+1l}}{k^2} + \dfrac{v_{ijl-1}-2v_{ijl}+v_{ijl+1}}{s^2} = -f_{ijl}$

$$A = -2\left(\frac{1}{h^2}+\frac{1}{k^2}+\frac{1}{s^2}\right)$$



Запишем шаблон для граничных узла: (0,j,l) (n,j,l) (i,0,l) (i,m,l) (i,j,0) (i,j,p).

Нетрудно видеть, что узлы, расположенные на ребрах куба в разностной схеме практически не используются – т.е. они не используются в уравнениях, аппроксимирующих основное уравнение задачи Дирихле. Таким образом, узлы, попадающие на ребра при изучении разностной схемы использовать не будем.

Сетка: H={h,k,s}

ω_H = топологически внутренние узлы (т.е. Ш'(i,j,l)≠0)

$$\omega_H = \{Ш'(i,j,l) \neq 0;\ i=1,n-1;\ j=1,m-1;\ l=1,p-1\}$$

γ_H – топологически граничные узлы:

$$\gamma_h = \{Ш'(i,j,l)=0;(i,0,l);(i,m,l);(0,j,l);(n,j,l),(i,j,0);(i,j,p);i=1,n-1;\ j=1,m-1;l=1,p-1\}$$

$$\Omega_H \overset{def}{=} \omega_H \cup \gamma_h$$

Сетка является связной.

Разностную схему (17) будем рассматривать на сетке (18), которая является связной. Запишем сетку (18) в каноническом виде:

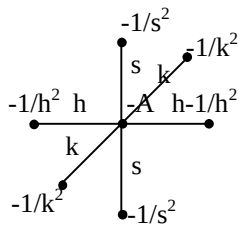
$$A(i,j,l)v_{ijl} - \sum_{(\hat{i},\hat{j},\hat{l})\in\Omega_H} B(i,j,l,\hat{i},\hat{j},\hat{l})v_{\hat{i}\hat{j}\hat{l}} = \Phi_{ijl},\ (i,j,l)\in\Omega_H \quad (19)$$ Для $(i,j,l)\in\omega_H$

Для того, чтобы разностная схема удовлетворяла принципу максимума, в основном уравнении (17*) поменяем знаки и нарисуем новый шаблон.

$$-A = 2\left(\frac{1}{h^2}+\frac{1}{k^2}+\frac{1}{s^2}\right) = A(i,j,l) > 0$$

$$B(i,j,l,i,j,l\pm1) = \frac{1}{s^2} \quad B(i,j,l,i\pm1,j,l) = \frac{1}{h^2}$$

$$B(i,j,l,i,j\pm1,l) = \frac{1}{k^2}$$



Условие теоремы о применении принципа максимума: $A(i,j,l)>0;\quad A-\sum B = 0 \ge 0$

Для $(i,j,l)\in\gamma_H$: $A(i,j,l)v_{ijl} = \Phi(i,j,l)$

<u>Итог</u>: $A(i,0,l)=1>0;\ \Phi(i,0,l)=\mu_{3il};\quad (i,j,l)\in\gamma_H;\ A(i,j,l)>0$

Разностная схема (17) удовлетворяет принципу максимума и имеет единственное решение.

<u>Теорема 3:</u>

Погрешность аппроксимации дифференциальной задачи (16) разностной схемы (17) удовлетворяет оценке: $$\max_{\substack{i=0,n\\ j=0,m\\ l=0,p}}|\psi_{ijl}| \le \frac{1}{12}\max_{(x,y,\omega)\in\bar{G}}\left\{\left|\frac{\partial^4 u}{\partial x^4}\right|,\left|\frac{\partial^4 u}{\partial y^4}\right|,\left|\frac{\partial^4 u}{\partial \omega^4}\right|\right\}\cdot\left(h^2+k^2+s^2\right) \quad (20)$$

<u>Доказательство:</u> аналогично двумерной задаче.

<u>Теорема 4:</u>

z_{ijl} и ψ_{ijl} связаны уравнением $L(i,j,l)z_{ijl}=-\psi_{ijl}$ для $(i,j,l)\in\omega_H$; $z_{ijl}=\psi_{ijl}=0$ для $(i,j,l)\in\gamma_H$

Здесь $$L(i,j,l)v_{ijl} = A(i,j,l)v_{ijl} - \sum_{(\hat{i},\hat{j},\hat{l})\in Ш'(i,j,l)} B(i,j,l,\hat{i},\hat{j},\hat{l})v_{\hat{i}\hat{j}\hat{l}} \quad (21)$$

Доказательство очевидно – доказать самим.

Наша цель – доказать сходимость разностной схемы (17) к решению задачи (16) – это можно сделать с помощью принципа максимума. Нужно построить мажорирующее уравнение (21).

Построим его с помощью $\hat{z}$:

$$(23)\ \hat{z}(x,y,\omega) = \frac{K}{6}\left(x(1-x)+y(1-y)+\omega(1-\omega)\right);\ K \overset{def}{=} M\left(h^2+k^2+s^2\right),\ \textit{где } M \textit{ из оценки } (20)$$

$$\begin{cases} \hat{z} = -K \\ \hat{z}\big|_{\partial G} = \ldots \ge 0 \end{cases}$$ - очевидное свойство

<u>Теорема 5:</u>

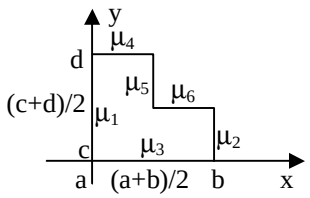
Если функция z и μ достаточно гладкие и точное решение задачи (16) достаточно гладкое, то решение разностной схемы (17) сходиться к решению дифференциальной задачи (16) равномерно по h, k и s с оценкой: $\max\limits_{\substack{i=0,n\\ j=0,m\\ l=0,p}} \left| z_{ijl} \right| \le \frac{l_1^2 + l_2^2 + l_3^2}{6 \cdot 4} \cdot M \left(h^2 + k^2 + s^2 \right),\ l_1 = l_2 = l_3 = 1 \quad (24)$

где $M = \frac{1}{12} \max\limits_{(x,y,\omega)\in\overline{G}} \left\{ \left| \frac{\partial^4 u}{\partial x^4} \right|, \left| \frac{\partial^4 u}{\partial y^4} \right|, \left| \frac{\partial^4 u}{\partial \omega^4} \right| \right\}$

Лекция №22.

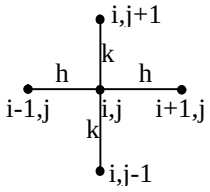
Задача Дирихле на прямоугольнике с выбитым квадрантом.

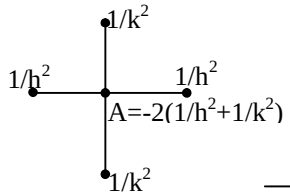
$$\begin{cases} u(x,y) = -f(x,y) & (x,y)\in G \\ u|_{\partial G} = \mu(x,y) \end{cases}$$

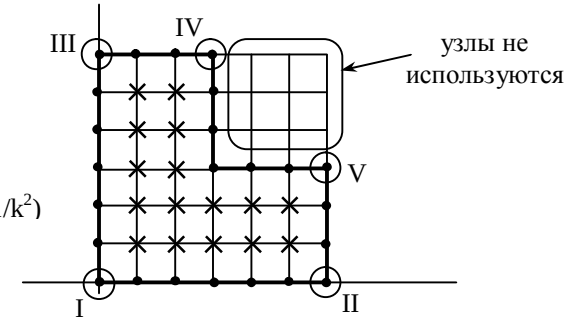


Функции μ согласованы в граничных точках.
Основа сетки:

$$(x_i, y_j): \quad \begin{aligned} x_i &= a + ih,\ i = 0,n \quad h = \frac{b-a}{n},\ n = 2\cdot n_1 \\ y_j &= c + jk,\ j = 0,m \quad k = \frac{d-c}{m},\ m = 2m_1 \end{aligned}$$







H={h,k}
топологически внутренние узлы – это те физически внутренние узлы, для которых можно применить шаблон крест × - ω_H.
γ_H – это те и только те граничные узлы, которые участвуют в шаблоне для внутренних узлов.
Физические граничные узлы I-V в множество γ_H не входят. Таким образом сеткой в нашей задаче считается $\omega_H \cup \gamma_H = \Omega_H$.
v_{ij} , $(i,j)\in\Omega_H$ – точное решение разностной схемы в узлах сетки.
u_{ij}, $(i,j)\in\Omega_H$ – точное решение дифференциальной задачи в узлах сетки.
$z_{ij}=u_{ij}-v_{ij}$, $(i,j)\in\Omega_H$ – погрешность.

$$(26)\begin{cases} \left(v_{x\bar{x}}\right)_{ij} + \left(v_{y\bar{y}}\right)_{ij} = -f_{ij} & (26*) \\ v_{ij} = \mu(x_i, y_j), (i,j)\in\omega_H \end{cases}$$

Введем оператор L(i,j) для $(i,j)\in\Omega_H$:

- для $(i,j)\in\omega_H$: $L(i,j)v_{ij} \overset{def}{=} -\left(\left(v_{x\bar{x}}\right)_{ij} + \left(v_{y\bar{y}}\right)_{ij}\right)$
- для $(i,j)\in\gamma_H$: $L(i,j)v_{ij} \overset{def}{=} v_{ij} \begin{cases} L(i,j)v_{ij} = f_{ij}, & (i,j)\in\omega_H \\ L(i,j)v_{ij} = \mu(x_i, y_j), & (i,j)\in\gamma_H \end{cases}$

Чтобы проверить удовлетворяет ли разностная схема (27) принципу максимума, рассмотрим общую запись для оператора L:

$$L(i,j)v_{ij} \overset{def}{=} A(i,j)v_{ij} - \sum_{(\hat{i},\hat{j})\in Ш'(i,j)} B(i,j,\hat{i},\hat{j})v_{\hat{i}\hat{j}}; \quad для\ (i,j)\in\omega_H : A(i,j) = 2\left(\frac{1}{h^2} + \frac{1}{k^2}\right) > 0$$

$$B(i,j,i\pm1,j) = \frac{1}{h^2} > 0; \quad B(i,j,i,j\pm1) = \frac{1}{k^2} > 0; \quad A - \sum B = 0 \ge 0;$$

$$для\ узлов\ (i,j)\in\gamma_H: \quad A(i,j) = 1 > 0$$

Теорема 6:

Разностная схема (27) удовлетворяет принципу максимума.

Теорема 7:

Для любой f, u из (25) решение разностной схемы (26) существует и единственно (следствие из принципа максимума).

Теорема 8:

Если f и μ из (25), а так же точное решение (25) u(x,y) достаточно гладкие, то решение разностной схемы (26) v_{ij}, $(i,j)\in\Omega_H$ сходиться к решению дифференциальной задачи равномерно со 2-м порядком по h и k:

$$(28)\ \max_{(i,j)\in\Omega_H}\left|z_{ij}\right| \le \frac{l_1^2+l_2^2}{4\cdot 4}M\left(h^2+k^2\right),\ \textit{где } M \textit{ из оценки для } \psi : M=\frac{1}{12}\max_{(x,y)\in\bar G}\left\{\left|\frac{\partial^4 u}{\partial x^4}\right|,\left|\frac{\partial^4 u}{\partial y^4}\right|\right\}$$

Доказательство: - самостоятельно.

Подсказки к доказательству теоремы 8:

1) Погрешность аппроксимации ψ_{ij}, $(i,j)\in\Omega_H$:

$$(i,j)\in\omega_H\ \psi_{ij}\overset{def}{=}\left(u_{x\bar x}\right)_{ij}+\left(u_{y\bar y}\right)_{ij}+f_{ij};\quad (i,j)\in\gamma_H\ \psi_{ij}\overset{def}{=}u_{ij}-\mu\left(x_i,y_j\right)=0$$

2) Для ψ и z нужно убедиться в том, что: $(29)\begin{cases} z_{ij}=\psi_{ij}=0 & (i,j)\in\gamma_H \\ \left(z_{x\bar x}\right)_{ij}+\left(z_{y\bar y}\right)_{ij}=\psi_{ij} & (i,j)\in\omega_H\end{cases}$

3) Убедиться в том, что (29) можно записать в виде (30):

$$(30)\begin{cases} L(i,j)z_{ij}=-\psi_{ij}, & (i,j)\in\omega_H \\ L(i,j)z_{ij}=\psi_{ij}, & (i,j)\in\gamma_H\end{cases}$$

4) Построить для (30) мажорирующее уравнение, используя:

$$\overset{B}{z}(x,y)=\frac{K}{4}\left((x-a)(b-x)+(y-c)(d-y)\right),\textit{где } K=M\left(h^2+k^2\right),\ \textit{где } M \textit{ из оц. для } \psi$$

5) Проверить, что $\overset{B}{z}$ удовлетворяет уравнению:

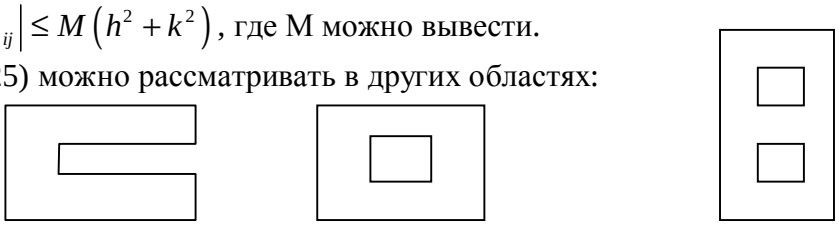
$$(32)\begin{cases} L(i,j)\overset{B}{z}_{ij}=K, & (i,j)\in\omega_H \\ L(i,j)\overset{B}{z}_{ij}=\underset{\textit{знач. } \overset{B}{z}(x,y) \textit{ на гран. мембраны}>0}{\ldots\ldots 1\ldots\ldots}, & (i,j)\in\gamma_H\end{cases}$$

Уравнение (32) мажорирует (30).

6) Из (30) и (32) следует, что $\left|z_{ij}\right|\le\overset{B}{z}_{ij}$, для $\forall(i,j)\in\Omega_H$.

7) В доказательстве использовалась оценка погрешности аппроксимации. $\left|\psi_{ij}\right|\le M\left(h^2+k^2\right)$, где M можно вывести.

Задачу (25) можно рассматривать в других областях:



и для них строить разностную схему, сходящуюся со вторым порядком.

Вспомогательная задача.

Рассмотрим f(x), сетка x_i, $x_{i-\alpha}=x_i-\alpha h$, $x_{i+\beta}=x_i+\beta h$.

α, β∈(0,1]

f”(x_i)-?

Строим интерполяционный полином $P_2(x)$ по 2-м точкам: $P_2(x_i)=f(x_i)$; $P_2(x_i+\beta h)=f(x_i+\beta h)$; $P_2(x_i-\alpha h)=f(x_i-\alpha h)$. Полагаем, что $P''_2(x_i)\sim f''(x_i)$.

Утверждение: $$P''_2(x_i)=\frac{2}{h^2}\left(\frac{1}{(\alpha+\beta)\alpha}f_{i-\alpha}-\frac{1}{\alpha\beta}f_i+\frac{1}{(\alpha+\beta)\beta}f_{i+\beta}\right)\quad(33)$$

причем: $$f''(x_i)=P_2(x_i)+\left(\frac{h(\alpha-\beta)}{3}f'''(x_i)-\frac{h}{12}\frac{\alpha^3+\beta^3}{\alpha+\beta}f^{IV}(x_i)+0\left(h^2\right)\right)\quad(34)$$

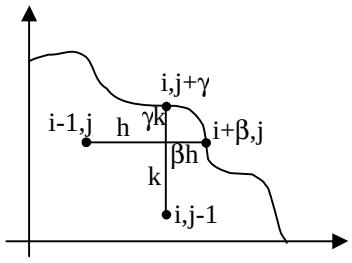
Доказательство: самостоятельно:
(33) – из дифференцирования полинома. (34) – через разложение в ряд Тейлора (33) в (.) x_i.

Определение:

Формулу (33) назовем 3-х точечным разностным оператором 2-й разносной производной на несимметричном шаблоне и обозначим $\left(f_{x\bar{x}}\right)_i = (33)$.

Свойства: 1) Если α=β, то $\left(f_{x\bar{x}}\right)_i = \frac{1}{(\alpha h)^2}\left(f_{i-\alpha} - 2f_i + f_{i+\alpha}\right)$, который аппроксимирует $f''(x_i)$ со 2-м порядком по h.

2) если α≠β, то $\left(f_{x\bar{x}}\right)_i$ аппроксимирует $f''(x_i)$ с 1-м порядком по h.

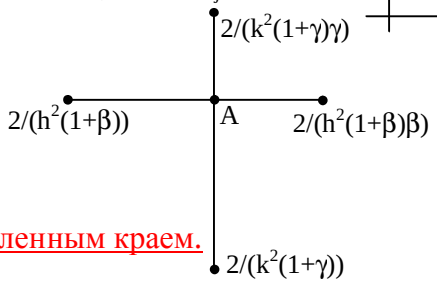


Утверждение:

На несимметричном шаблоне, непрямоугольном сетке u в узле i,j аппроксимируется разностным оператором $\left(u_{x\bar{x}}\right)_{ij} + \left(u_{y\bar{y}}\right)_{ij}$ с первым порядком по h и k.

$$A = -\frac{2}{h^2\beta} - \frac{2}{k^2\gamma} = -2\left(\frac{1}{h^2\beta} + \frac{1}{k^2\gamma}\right)$$

при β=γ=1 получим шаблон крест.



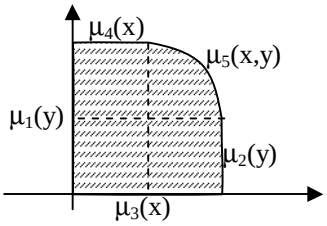
Задача Дирихле на прямоугольнике с закругленным краем.

Основа сетки:

$$x_i = a + ih \quad h = \frac{b-a}{n}, \quad i = 0, n$$

$$y_j = c + jk \quad k = \frac{d-c}{m}, \quad j = 0, m$$

$$n = 2n_1 \quad m = 2m_1$$



ω_H – те внутренние узлы, на которых можно использовать либо симметричный шаблон, либо несимметричный.

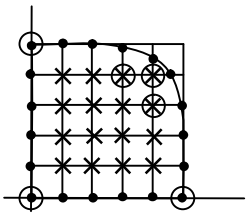
γ_H – физические граничные узлы, которые участвуют в уравнении для узлов из ω_H.

Сетка: $\Omega_H = \gamma_H \cup \omega_H$, H={h,k}

Узлы I-III в γ_H не входят.

Разностная схема:

$$(37)\begin{cases} \left(v_{x\bar{x}}\right)_{ij} + \left(v_{y\bar{y}}\right)_{ij} = -f_{ij}, \ (i,j) \in \omega_H \\ v_{ij} = \mu\left(x_i, y_j\right), \ (i,j) \in \gamma_H \end{cases}$$



Разностная схема (37) удовлетворяет принципу максимума ⇒

- существует единственное решение разностной схемы
- есть сходимость к решению дифференциальной задачи с 1-м порядком по h и k.

Лекция №23.

Мы умеем строить разностную схему 1,2-го порядка аппроксимации для области любой размерности, если ее можно заключить в многомерный параллелепипед.

6.5 Решение разностных схем для настоящих задач.

На примере двумерных задач. Составим вектор $\bar{v}$, компоненты которого соответствуют значениям сеточной функции в топологически граничных узлах. При этом заранее обговаривают правило обхода.

Уравнение разностной схемы, соответствующей топологически внутренним узлам записывают в виде матрицы: $A\bar{v} = F$.

Для нумерации компонент вектора строк и столбцов матрицы удобно оставит двойную индексацию: v_{ij} – компонента вектора $\bar{v}$, соответствующая узлу (x_i,y_j); (i,j) – строка матрицы A, которая соответствуют уравнению, ассоциированному с узлом (x_i,y_j); (l,s) – столбец матрицы A. Элемент матрицы A в строке (i,j), столбце (l,s) равен коэффициенту с которым v_{ls} входит в уравнение, ассоциированное с узлом (x_i,y_j). Вектор F содержит слагаемые правой части уравнения, а так же значения функции в топологически – граничных узлах.

Проверка симметричности:

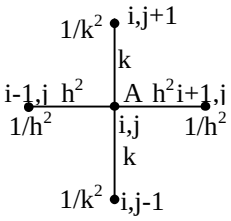
- в строке (i,j) столбце (l,s) участие (l,s) в уравнении для (i,j).
- в строке (), столбце (i,j) участие (i,j) в уравнении для (l,s).

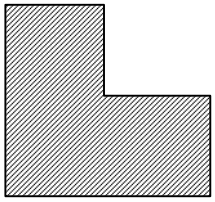
Утверждение:

Матрица разностной схемы задачи Дирихле симметрична.

Доказательство:

$$A = -2\left(\frac{1}{h^2} + \frac{1}{k^2}\right)$$





Очевидно, что участие коэффициентов i±1,j в уравнение для i,j будет таким же, как и участие (i,j) для уравнений (i±1,j).

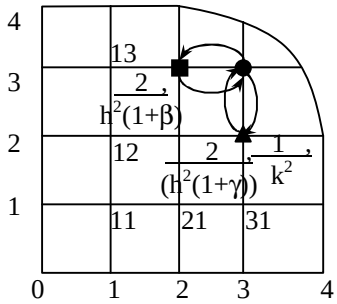
Утверждение:

Разностная схема, для решения задачи Дирихле в 3-х мерном кубе симметрична: $A_{3D}=A^T_{3D}$.

Утверждение:

Матрица разностной схемы, для решения задачи Дирихле в области с закругленным краем не симметрична. $A_0 \neq A^T_0$.

Доказательство: Узел (3,3) участвует в уравнении для (2,3) с коэффициентом $1/h^2$, а узел (2,3) участвует в уравнении для (3,3) с коэффициентом $\frac{2}{h^2(1+\beta)}$.



Узел (3,3) участвует в уравнении для (3,2) с коэффициентом $\frac{1}{k^2}$, а (3,2) участвует в уравнении для (3,3) с коэффициентом $\frac{1}{k^2(1+\gamma)}$

Проверка невырожденности: $\det A \neq 0$ – это следует из принципа максимума.

Проверка расположения собственных чисел – по теореме Гершгорина.

При подборе методов решения разностных схем мы пользуемся свойством:

1) Если $A=A^T$, то все ее собственные числа действительные и существует ортонормированный базис из собственных векторов.
2) Если $A=A^T$, то следующие 2 утверждения эквиваленты:
   a. все $\lambda_i(A)>0$, i=1,n
   b. A>0.

   Доказательство: $h \in R^n$, $h \neq 0$; $v_1,\ldots,v_n$ – ортонормированный базис из собственных векторов.

   $h = \sum \alpha_i v_i$, причем $\exists l$, $\alpha_l \neq 0$.

   Рассмотрим $(Ah,h) = \left(A\sum_{i=1}^{n}\alpha_i v_i, \sum_{s=1}^{n}\alpha_s v_s\right) = \left(\sum_{i=1}^{n}\alpha_i\lambda_i v_i, \sum_{s=1}^{n}\alpha_s v_s\right) = \sum_{i=1}^{n}\underset{\geq 0}{\alpha_i}\,\underset{>0}{\lambda_i}\,\underset{>0}{(v_i,v_i)} \geq \underset{>0}{\alpha_l^2}\,\underset{>0}{\lambda_l}\,\underset{>0}{(v_l,v_l)}$

   оставим только ненулевое слагаемое.

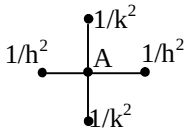
   то $\forall h \neq 0$ $(Ah,h)>0 \Rightarrow (A>0)$

Пусть A>0, λ, v – собственная пара A; $v \neq 0$ $(Av,v)=\lambda(v,v)>0 \Rightarrow \lambda>0$.

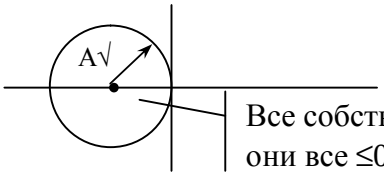
Если матрица симметрична, то без разницы что проверять – положительную определенность матрицы или собственных чисел.

Многие итерационные методы алгебры предполагают, что матрица: $A=A^T>0$. Используя свойство №2 будем проверять симметричность и расположение собственных чисел матриц разностных схем.

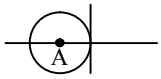
1)  В строке матрицы $A_{:.}$ ассоциированной с (i,j) используется шаблон:



Круг Гершгорина:



Все собственные числа в круге и действительные ⇒ они все ≤0.

- Если узел (i,j) является узлом 1-го типа, то круг Гершгорина имеет вид выше на рисунке.
- Если узел (i,j) является узлом 2-го типа, то  Все собственные числа лежат в кругах Гершгорина или на их границе, собственные числа симметричной матрицы действительны, так как выполнятся принцип максимума – нулевого собственного числа нет → Все собственные числа $A_{:.}$ отрицательны.

Теорема:

Систему $-A_{..}\bar{v} = F$ можно решать итерационными методами.

Замечания: если разностная схема удовлетворяет принципу максимума, то круги Гершгорина не будут пересекать мнимую ось.

Теорема:

Разностную схему 3-х мерной задачи после замены знака можно решать теми же методами, т.к. матрица $A_{3D} = A_{3D}^T > 0$: $-A_{3D}\bar{v} = -F$

Доказательство – аналогично.

Теорема:

Матрица разностной схемы для закругленной области не симметрична, но свойство диагонального преобладания имеется.

Доказательство: Используется один из шаблонов:

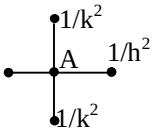


рис. 1

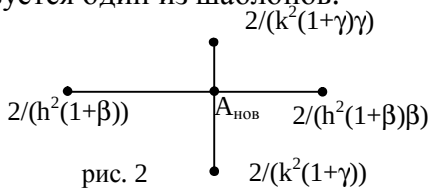


рис. 2

$$A_{нов} = -2\left(\frac{1}{h^2\beta} + \frac{1}{k^2\gamma}\right)$$

В строках матрицы A•, использующих шаблон рисунка 1 $|A| \ge \left|\frac{1}{h^2}\right| + \left|\frac{1}{h^2}\right| + \left|\frac{1}{k^2}\right| + \left|\frac{1}{k^2}\right|$.

В строках матрицы $A_{нов}$, использующих шаблон на рисунке 2

$$A_{нов} \ge \frac{2}{h^2(1+\beta)}\left(1+\frac{1}{\beta}\right) + \frac{2}{k^2(1+\gamma)}\left(1+\frac{1}{\gamma}\right)$$

#7 Линейная Алгебра.

#7.1 Норма матриц и векторов.

$x=(x^{(1)},\ldots,x^{(n)}) \in R^{(n)}$

Определение:

Нормой вектора ||x|| называется функционал, удовлетворяющий 3-м аксиомам нормы:

1) $(\|x\| = 0) \Leftrightarrow (x = 0)$ и $\|x\| > 0$

2) $\|ax\| = |a| \cdot \|x\|$

3) $\|x + y\| \le \|x\| + \|y\|$

Пример: $\|x\|_2 \overset{def}{=} \sqrt{(x,x)}$ – *евклидова норма*; $\|x\|_1 \overset{def}{=} \sum_{i=1}^{n} \left|x^{(i)}\right|$; $\|x\|_\infty \overset{def}{=} \max_{i=1,n} \left|x^{(i)}\right|$

Определение:

Нормой матрицы называется функционал ||A||, удовлетворяющий 4 аксиомам нормы:
1) $\|A\| = 0 \Leftrightarrow A = 0,\ \|A\| \ge 0$; 2) $\|\alpha A\| = |\alpha| \cdot \|A\|$; 3) $\|A+B\| \le \|A\| + \|B\|$; 4) $\|A \cdot B\| \le \|A\| \cdot \|B\|$

Определение:

$\|A\| = \sup_{x \ne 0} \frac{\|Ax\|}{\|x\|}$ – норма $\|A\|$, подчиненная норме ||x||.

Пример: $\|A\|_2 = \sup_{x \ne 0} \frac{\|Ax\|_2}{\|x\|_2}$

Лекция №24.

Без ограничения общности подчиненную норму можно определить: $\|A\| \overset{def}{=} \sup_{\|x\|=1} \|Ax\|$ (1*)

Доказательство: $\|A(\alpha x)\| = |\alpha| \cdot \|Ax\|$, $\|\alpha x\| = |\alpha| \cdot \|x\|$

В численных методах линейной алгебры нужны оценки вида ||Ax||≤M·||x|| (2) верные для любого x.
С помощью подчиненных норм можно строить точные оценки вида (2).

Утверждение 1:

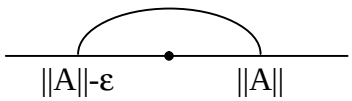
Если матричная норма подчинена векторной, то ∀A(n,n), ∀x∈ R^n выполнена оценка: ||Ax||≤||A||·||x||.

Доказательство: x=0, 0≤0. x≠0, $\|Ax\| = \frac{\|Ax\|}{\|x\|} \cdot \|x\| \le \underbrace{\sup_{y \ne 0} \frac{\|Ay\|}{\|y\|}}_{\|A\|} \|x\|$

Утверждение 2:

Пусть дана A(n×n), A≠0, тогда оценки вида $\|Ax\| \le M\|x\|$ (**), где 0<M<||A|| не могут быть верными для любого x.
(т.е. оценка ||Ax||≤||A||·||x|| не улучшаема)

Доказательство: sup: ∀ε>0 ∃x*; что $\underbrace{\sup_{x \ne 0} \frac{\|Ax\|}{\|x\|}}_{\|A\|} - \varepsilon < \frac{\|Ax^*\|}{\|x^*\|} \le \underbrace{\sup_{x \ne 0} \frac{\|Ax\|}{\|x\|}}_{\|A\|}$



Пусть для некоторого A≠0 существует M, 0<M<||A||, что верна оценка (**) для любого x. Тогда ε=||A||-M – подставим в (*).

$\Rightarrow M < \frac{\|Ax^*\|}{\|x^*\|} \le \|A\|$, $\|Ax^*\| > M\|x^*\| \Rightarrow$ оценка (**) не верна для любого x.

Вычисление матричных норм.

Определение:

Набор собственных чисел матрицы A называется спектром.

Определение:

Спектральным радиусом матрицы A называют расстояние от 0 комплексной плоскости до самой далекой точки спектра. $p(A) = \max_{i=1,n} \{|\lambda_i(A)|\}$ (3)

Теорема 1: $\|A\|_2 = \sqrt{p(A^T, A)}$ (4) - матричная норма, подчиненная Евклидовой норме векторов.

Теорема 2: $\|A\|_1 = \max_{j=1,n} \sum_{i=1}^{n} |a_{ij}|$ (5) – подчиненная норме $\|x\|_1$.

Теорема 3: $\|A\|_\infty = \max\limits_{i=1,n} \sum\limits_{j=1}^{n} |a_{ij}|$ (6) - подчиненная норме $\|x\|_\infty$.

Пример:
$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \begin{matrix} \to 6 \\ \to 15 \\ \to 24 \end{matrix} \qquad \begin{matrix} \|A\|_1 = 24 \\ \|A\|_\infty = 18 \end{matrix}$$
$$\begin{matrix} \downarrow & \downarrow & \downarrow \\ 12 & 15 & 18 \end{matrix}$$

Формулы (5) и (6) хороши тем, что для расчета нормы матрицы не нужны собственные числа.

Есть теорема, позволяющая оценить $\|A\|_2$ через $\|A\|_1$ и $\|A\|_\infty$.

Утверждение:

В конечномерном (Гильбертовом) пространстве все векторные нормы эквивалентны. Если некоторая последовательность векторов сходится к 0 в одной норме, то она сходится к 0 и в другой.

Вывод при доказательстве сходимости итерационных методов линейной алгебры можно использовать любые матричные нормы и любые нормы векторов.

7.2 Собственная пара.

Определение:

Число λ и вектор $v \neq 0 \in R^n$ называются собственной парой матрицы $A(n\times n)$, если $Av=\lambda v$.
v - собственный вектор; λ - собственное число.

Утверждение:

Собственные числа матрицы A являются решениями характеристического уравнения $\det(A-\lambda E)=0$ (8), причем каждому собственному числу соответствует по крайней мере 1 собственный вектор (с точностью до линейной независимости).

Определение:

Алгебраическая кратность λ - кратность числа корней в уравнении (8).

Определение:

Геометрической кратностью λ называется количество линейно независимых собственных векторов, соответствующих числу λ.

Утверждение: $1 \le$ геометрическая кратность $\le$ алгебраическая кратность.

Из-за того, что геометрическая кратность может быть меньше алгебраической, не все матрицы имеют базис из собственных векторов.

Утверждение:

Симметричная матрица $A=A^T$ имеет ортонормированный базис из собственных векторов.

Утверждение:

(λ, v) – собственная пара матрицы A $\Rightarrow$ а) (λ^2, v) – собственная пара A^2; б) $(1/\lambda, v)$ – собственная пара A^{-1}; в) $(1-\tau\lambda, v)$ – собственная пара $E-\tau A$;
г) $\left((1-\tau_1\lambda)(1-\tau_2\lambda)\ldots(1-\tau_k\lambda), v\right)$ – собственная пара $(E-\tau_1 A)(E-\tau_2 A)\ldots(E-\tau_k A) = B$

Доказательство: а) $A^2v=Aav=A\lambda v=\lambda^2 v$
б) $Av=\lambda v \Rightarrow A^{-1}Av=A^{-1}(\lambda v) \Rightarrow v=\lambda A^{-1}v \Rightarrow A^{-1}v=1/\lambda v$, $\lambda \neq 0$.
в) $(E-\tau A)v=v-\tau\lambda v=(1-\lambda\tau)v$
г) рассмотрим B и (λ,v): $Bv = (E-\tau_1 A)\ldots(E-\tau_k A)v = (E-\tau_1 A)\ldots(E-\tau_{k-1} A)(1-\tau_k\lambda)v = \ldots$

Спектральный радиус обратной матрицы.

Теорема 4: пусть $A(n\times n)$, $\det A \neq 0 \Rightarrow$ $p(A^{-1}) = \dfrac{1}{\min\limits_{i=1,n}\{|\lambda_i(A)|\}}$

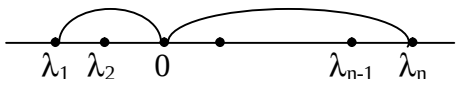
Доказательство: $p(A)=\max\{|\lambda_i(A)|\}$, $\exists A^{-1}$. Если (λ, v) – собственная пара матрицы A, то (1/λ, v) – собственная пара A^{-1} ⇒ если $(\lambda, \vec{v})$ – собственная пара A^{-1} ⇒ $(1/\lambda, \vec{v})$ - собственная пара A.

$$\Rightarrow \underset{i=1,n}{\lambda_i(A)} - \text{спектр A} \leftrightarrow \underset{i=1,n}{1/\lambda_i(A)} - \text{спектр } A^{-1}.\ p(A^{-1}) \overset{def}{=} \max_{i=1,n}\left\{\left|\frac{1}{\lambda_i(A)}\right|\right\} = \frac{1}{\min_{i=1,n}\{|\lambda_i(A)|\}}.$$

7.3 Взаимосвязь спектра и нормы матрицы.

Утверждение:

Если $A=A^T$ ⇒ все собственные числа действительны и спектр можно упорядочить: $\lambda_1 \le \lambda_2 \le \ldots \le \lambda_n$ причем $p(A)=\max\{|\lambda_1|,|\lambda_n|\}$



Утверждение: (*) Если $A=A^T$, то $\|A\|_2 = p(A)$ (10)

Утверждение (**):

Если $A=A^T$, то любая подчиненная норма $\|A\| \ge p(A)$ (11).

Утверждение (***):

Для любой A и для любой подчиненной нормы $\|A\| \ge p(A)$ (12).

Доказательство (**): пусть (λ, v) собственная пара A, такая, что спектральный радиус $p(A)=|\lambda|$.

Тогда: $\frac{\|Av\|}{\|v\|} = \frac{|\lambda|\cdot\|v\|}{\|v\|} = |\lambda| \le \sup_{y\ne 0}\frac{\|Ay\|}{\|y\|} \Rightarrow p(A) \le \|A\|$

Доказательство (*): $A^TA=A^2$; если (λ, v) – собственная пара A, то (λ^2, v) – собственная пара A^2.

$$p(A^TA) = p(A^2) = \max_{i=1,n}\left\{|\lambda_i(A)|^2\right\} = \max\{\lambda_1^2,\lambda_n^2\};\ \ p(A)=\max\{|\lambda_1|,|\lambda_n|\}$$

$$\|A\|_2 = \sqrt{\max\{\lambda_1^2,\lambda_n^2\}} = \max\{|\lambda_1|,|\lambda_n|\} = p(A)$$

(***) доказывается так же, как (**). Нужно взять λ, такой, на котором достигается спектральный радиус и рассмотреть 2 случая – когда он действительный и комплексный.

Лекция №25. Инструменты для изучения сходимости.

7.4 Число обусловленности.

Определение:

A(n×n): $\det A \ne 0$ ⇒ числом обусловленности называется $\mu_A = \|A\|\cdot\|A^{-1}\|$ (1)

Определение:

$\det A \ne 0$ ⇒ числом обусловленности Тодта называется : $\mu_A = \frac{\max_{i=1,n}\{|\lambda_i(A)|\}}{\min_{i=1,n}\{|\lambda_i(A)|\}}$ (2)

Утверждение 1:

Пусть $\det A \ne 0$ ⇒ для любого способа задание вектора нормы и соответственно любой подчиненной матричной нормы $\mu_A \ge \mu_A^T \ge 1$ (3)

Доказательство: для любой подчиненной матричной нормы $\|A\| \ge p(A) = \max_{i=1,n}\{|\lambda_i(A)|\}$ и

$$\|A^{-1}\| \ge p(A^{-1}) = 1/\min_{i=1,n}\{\lambda_i(A)\} \Rightarrow \mu_A \ge \mu_A^T \ge 1, т.к.\ \frac{\max}{\min} \ge 1$$

Утверждение 2: пусть $A=A^T>0$ ⇒ $\mu_A^T = \frac{\lambda_n}{\lambda_1}$, где $0<\lambda_1 \le \lambda_2 \le \ldots \le \lambda_n$ собственные числа A

Утверждение 3: пусть $A=A^T>0$ ⇒

1) В любой норме $\mu_A \ge \lambda_n/\lambda_1 \ge 1$

2) Если для вектора используется евклидова норма, а для матрицы норма, подчиненная евклидовой $\Rightarrow \mu_A = \mu_A^T = \lambda_n / \lambda_1$

Доказательство: $\left.\begin{array}{l} \|A\|_2 = p(A) = \lambda_n \\ \|A^{-1}\|_2 = p(A^{-1}) = 1/\lambda_n \end{array}\right\}$ см. в пред. лекции $\Rightarrow \mu_A = \mu_A^T = \lambda_n / \lambda_1$

Замечание 1: Если $A=A^T>0 \Rightarrow \mu_A = \mu_A^T = \dfrac{\lambda_n}{\lambda_1} \geq 1 \; (4)$ в евклидовой норме.

Замечание 2: Число обусловленности играет важную роль при изучении сходимости итерационных методов линейной алгебры, а так же при изучении реакции линейной системы на возмущение правой части (#7.9, #7.5)

7.5 Обусловленность систем линейных уравнений.

Рассмотри 2 системы: Ax=b, ее точное решение x* (5)

$$Ax = b + \;\; b, ee \text{ точное решение } x^* + \;\; x \; (6)$$

detA≠0 b – возмущение правой части системы (6), x – возмущение решения системы (6).

(7) $\dfrac{\|\;\; b\|}{\|b\|}$ – относительная погрешность возмущения правой части.

(8) $\dfrac{\|\;\; x\|}{\|x^*\|}$ - относительное возмущение решения.

Замечание: В случае b=0 ⇒ ||b||=0 мы его не рассматриваем, т.к. исследователь должен заранее знать, что в правой части системы чистый 0.

$$\|b\| = \|Ax^*\| = \|A\| \cdot \|x^*\| \Rightarrow \|x^*\| \geq \frac{\|b\|}{\|A\|} \qquad A(x^* + \;\; x) = b + \;\; b \qquad A \;\; x = \;\; b$$

$$x = A^{-1} \;\; b, \; \|\;\; x\| = \|A^{-1} \;\; b\| \leq \|A^{-1}\| \cdot \|b\| \qquad \frac{\|\;\; x\|}{\|x^*\|} \leq \frac{\|A^{-1}\| \cdot \|\;\; b\| \cdot \|A\|}{\|b\|}$$

$$\boxed{\frac{\|\;\; x\|}{\|x^*\|} \leq \mu_A \frac{\|\;\; b\|}{\|b\|} \quad (9)}$$

Утверждение 4:

Для любой A, такой, что detA≠0 и любого способа выбора вектора и подчиненной матричной нормы верна оценка (9).

(9) – показывает, как отреагирует система линейных уравнений на возмущение правой части.

Определение:

Система линейных уравнений называется плохо обусловленной, если в какой-либо из норм число μ_A – большое.

Определение:

Система линейных уравнений называется хорошо обусловленной, если в какой-либо из норм $\mu_A \approx 1$.

Замечание: 1) Чем меньше μ_A, тем спокойнее реагирует система на возмущение правой чисти.

2) Оценки (9) улучшить нельзя, т.е. для любой матрицы A: detA≠0 существует такая правая часть b и существует такое b, что (9) выполнена «почти как равенство, с любой степенью точности».

т.к. (9) строилась на основе подчиненных норм, обеспечивающих наиболее точную оценку вида:

$$\|Ax\| \leq M \cdot \|x\| \quad (10)$$

Таким образом, плохо обусловленная система может «бешено» отреагировать на маленькое возмущение правой части.

7.6 Механизм обусловленности (при $A^T=A>0$)

Рассмотрим $A=A^T>0$, $0<\lambda_1\le\lambda_2\le\ldots\le\lambda_n$, пусть $\lambda_n>>\lambda_1$

$v_1,\ldots,v_n$ – собственные векторы (будем их считать ортонормированным базисом).

Для любого вектора $h\in R^n$ есть разложение $h=\sum_{i=1}^{n}\alpha_i v_i=\alpha_1 v_1+\ldots+\alpha_n v_n$

$$Ah=\sum_{i=1}^{n}\alpha_i Av_i=\lambda_1\alpha_1 v_1+\ldots+\lambda_n\alpha_n v_n$$

Утверждение:

Если h будет похож на v_1 и v_n одинаково, то Ah будет больше похож на v_n чем на v_1.

Относительный вклад v_n после применения матрицы A будет больше, чем относительный вклад v_1.

Используем тот же прием для изучения решения системы Ax=b (11).

$b=\beta_1 v_1+\ldots+\beta_n v_n$ – разложим по базису. x* - точное решение системы.

$$x^*=\frac{\beta_1}{\lambda_1}v_1+\frac{\beta_2}{\lambda_2}v_2+\ldots+\frac{\beta_n}{\lambda_n}v_n$$

Утверждение:

Если правая часть системы (11) была одинаково похожа на v_1 и v_n, то x* больше похож на v_1, чем на v_n.

Если правые чисти системы одинаково похожи на ортонормированные собственные вектора, то решение системы больше похоже на собственные векторы, соответствующие меньшим собственным числам.

Пример: когда оценка (9) выполнятся как равенство: $A=A^T$, $0<\lambda_1\le\lambda_2\le\ldots\le\lambda_n$, $\lambda_n>>\lambda_1$.

$\|v_1\|=\ldots=\|v_n\|$, $\alpha\approx 0-$ *число*

Рассмотрим 2 системы:

$$Ax=v_n\ (5^*)\ \text{малое возмущение}$$

$$Ax=v_n+\underset{\downarrow}{\alpha v_1}\ (6^*)$$

(6*) – система (5*) с возмущенной правой частью.

$$x^*=\frac{v_n}{\lambda_n};\quad x^*+\ x=\frac{v_n}{\lambda_n}+\alpha\frac{v_1}{\lambda_1};\quad b=\alpha v_1;\ b=v_n;\quad x=\frac{\alpha v_1}{\lambda_1}$$

$$\frac{\|x\|}{\|x^*\|}=\frac{|\alpha|\cdot\|v_1\|\lambda_n}{\lambda_1\|v_n\|}=|\alpha|\cdot\frac{\lambda_n}{\lambda_1}=\underbrace{\frac{\lambda_n}{\lambda_1}}_{\mu_A}\frac{\|b\|}{\|b\|};\qquad \frac{\|b\|}{\|b\|}=\frac{|\alpha|\cdot\|v_1\|}{\|v_n\|}=|\alpha|$$

Утверждение:

Пусть $A=A^T>0$ и $0<\lambda_1\le\lambda_2\le\ldots\le\lambda_n$. Пусть b пропорционален собственному вектору, соответствующему наибольшему собственному числу и пусть b пропорционален собственному вектору, соответствующему собственному числу.

$\Rightarrow$ в евклидовой норме оценка (9) выполняется как равенство: $\frac{\|x\|}{\|x^*\|}=\mu_A\cdot\frac{\|b\|}{\|b\|}$ (12) (если $\lambda_n>>\lambda_1$ – это плохо).

Пример: $-A\bar{v}=-F$ из разностной схемы задачи Дирихле (n,m).

7.7 Метод простой итерации (МПИ).

Ax=b (13), $\det A\ne 0$; x* - точное решение. $x_0\in R^n$– начальное приближение. τ - параметр метода.

$$\frac{x_{s+1}-x_s}{\tau}+Ax_s=b$$

Выяснить каким должен быть τ, чтобы метод сходился. Изучим переходную матрицу:

$x_{s+1}=(b-Ax_s)\tau+x_s$; $z_s=x_s-x^*$ - погрешность для x_s.

$z_{s+1} = x_{s+1} - x_s = \underbrace{(b - Ax_s)}_{-\tau A z_s}\tau + \underbrace{x_s - x_s}_{z_s}$ переходная матрица G=E-τA.

$z_{s+1}=Gz_s=G^2z_{s-1}=\ldots=G^{s+1}z_0$, где $z_0=x_0-x^*$ - погрешность начального приближения.

Утверждение: О достаточных условиях сходимости МПИ:

$$\|G\|\le 1$$

Доказательство: $\|z_{s+1}\| \le \|G\|\cdot\|z_s\| \le \|G\|^2\|z_{s-1}\| \le \ldots \le \|G\|^{s+1}\|z_0\|; \to \|z_{s+1}\| \le \underbrace{\|G\|^{s+1}}_{\to 0, s\to\infty}\|z_0\| \to 0$

Утверждение: спектральный радиус переходной матрицы $p(G)<1$- необходимое условие сходимости МПИ.

Доказательство: самим.

7.8 Метод простой итерации для A^T=A>0.

$0<\lambda_1\le\lambda_2\le\ldots\le\lambda_n$

если (λ,v) – собственная пара матрицы A ⇒ (1-τλ, v) – собственная пара G=E-τA.

если A^T=A>0, то $G^T = G \Rightarrow \|G\|_2 = p(G) = \max\limits_{i=1,n}\{|1-\tau\lambda_i|\}$

Используем достаточное условие сходимости:

$$\|G\|_2 < 1 \Rightarrow |1-\lambda_i\tau| < 1 \sim -1 < 1-\lambda_i\cdot\tau < 1,\ i=1,n$$

$$0<\tau\lambda_i<2,\ 0<\tau<\frac{2}{\lambda_i},\ \ i=1,n$$

⇒ τ∈(0, 2/λ_n) – достаточное условие сходимости метода.

Получим оценку скорости сходимости: т.к. τ>0 ⇒ собственные числа можно упорядочить:

$1 > \underbrace{1-\tau\lambda_1 \ge 1-\tau\lambda_2 \ge \ldots \ge 1-\tau\lambda_n}_{\text{собственные числа } G} \to \|G\|_2 = \max\{|1-\tau\lambda_1|, |1-\tau\lambda_n|\}$

Утверждение:

Если A=A^T>0 ⇒ МПИ сходиться к решению системы Ax=b из любого начального приближения x_0, при любом τ∈(0, 2/λ_n) с оценкой: $\|z_{s+1}\| \le \left(\max\{|1-\lambda_1\tau|, |1-\lambda_n\tau|\}\right)^{s+1}\|z_0\|_2$

Лекция №26.

Полученные оценки позволяют оценить $\|x_{s+1}-x^*\|$ *через* $\|x_s - x^*\|$

#7.9 МПИ с оптимальными параметрами.

A=A^T>0, $0<\lambda_1\le\lambda_2\le\ldots\le\lambda_n$

$$\frac{x_{s+1}-x_s}{\tau} + Ax_s = b \quad (14^*)$$

Если τ∈(0, 2/λ_n), то МПИ сходится с оценкой $\|x_{s+1}-x^*\|_2 \le \underbrace{\left(\max\{|1-\lambda_1\tau|, |1-\lambda_n\tau|\}\right)}_{F(\tau)}^{s+1}\|x_0 - x^*\|_2$

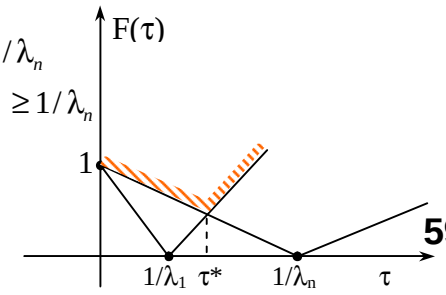
$$\Rightarrow \|x_{s+1}-x^*\|_2 \le \left(F(\tau)\right)^{s+1}\cdot\|x_0-x^*\|_2$$

Подберем такое τ*, для которого гарантированна наиболее быстрая сходимость.

$$\boxed{F(\tau)\to\min,\ где\ \tau\in\left(0,\frac{2}{\lambda_n}\right)}\ (15)\ \text{т.е.}\ \min_{\tau\in(0,2/\lambda_n)}\left(\max\{|1-\lambda_1\tau|, |1-\lambda_n\tau|\}\right)$$

$$|1-\lambda_1\tau| = \begin{cases}1-\lambda_1\tau,\ если\ \tau<1/\lambda_1\\ -(1-\lambda_1\tau),\ если\ \tau\ge 1/\lambda_1\end{cases} \qquad |1-\lambda_n\tau| = \begin{cases}1-\lambda_n\tau,\ если\ \tau<1/\lambda_n\\ -(1-\lambda_n\tau),\ если\ \tau\ge 1/\lambda_n\end{cases}$$

|*|=0, если τ=1/λ_1 или τ=1/λ_n, соответственно. |*|=1, если τ=0.



Очевидно, что график $F(\tau)$ можно построить по 3-м точкам и минимум находиться в τ^*, найдем его:

$$1-\lambda_1\tau=-(1-\lambda_n\tau) \;\rightarrow\; 2=(\lambda_1+\lambda_n)\tau^* \Rightarrow \tau^* = \frac{2}{(\lambda_1+\lambda_n)} \in \left(0, \frac{2}{\lambda_n}\right)$$

Построим оценку сходимости для МПИ с параметром τ^*:

$$F(\tau^*) = 1-\lambda_1\tau^* = -(1-\lambda_n\tau^*) = 1-\lambda_1\left(\frac{2}{\lambda_1+\lambda_n}\right) = \ldots = \frac{\lambda_n-\lambda_1}{\lambda_n+\lambda_1} = \frac{\frac{\lambda_n}{\lambda_1}-1}{\frac{\lambda_n}{\lambda 1}+1}$$

Теорема:

Если $A=A^T>0$, то наилучшую оценку сходимости МПИ для решения системы $Ax=b$, дает параметр

$$\tau^* = \frac{2}{\lambda_1+\lambda_n}, \; \textit{причем} \; \|x_{s+1}-x^*\|_2 \le \left(\frac{\frac{\lambda_n}{\lambda_1}-1}{\frac{\lambda_n}{\lambda 1}+1}\right)^{s+1} \cdot \|x_0-x^*\|_2$$

$$\mu_A = \frac{\lambda_n}{\lambda_1} - \text{число обусловленности} \;\Rightarrow \|z_{s+1}\|_2 \le \left(\frac{\mu_A-1}{\mu_A+1}\right)^{s+1} \cdot \|z_0\|_2 \quad (16^*)$$

Следствие:

Если матрица будет плохо обусловлена, то сходимость будет медленной, т.е. при $\lambda_n >> \lambda_1$.

7.1 Метод минимальных невязок.

(1) $Ax=b$, $x\in R^n$, $\det A\neq 0$; x^* - точное решение.

(2) рассмотрим $\dfrac{x_{s+1}-x_s}{\tau_s} + Ax_s = b$

τ_s – параметр для вычисления x_{s+1}. x_0 – начальное приближение.

Выясним, как в ходе работы метода меняется погрешность z_s: $z_s \overset{def}{=} x_s - x^*$, $r_s \overset{def}{=} Ax_s - b$

$$x_{s+1} = (b-Ax_s)\tau_s + x_s \;\Big|-x^* \rightarrow \underbrace{x_{s+1}-x^*}_{z_{s+1}} = \underbrace{(b-Ax_s)}_{\tau_s A z_s}\tau_s + \underbrace{x_s-x^*}_{z_s} \Big| \cdot A \textit{ слева}$$

$$\underbrace{A(x_{s+1}-x^*)}_{r_{s+1}} = \underbrace{A(b-Ax_s)}_{-r_s}\tau_s + \underbrace{A(x_s-x^*)}_{r_s} \Rightarrow \begin{matrix} r_{s+1} = (E-\tau_s A)r_s \\ (E-\tau_s A)z_s \end{matrix} \quad (3)$$

Во всех методах и системах: $\boxed{r_s = Az_s}$ (3^*)

Доказательство: $A(x_s-x^*)=Ax_s-b=r_s$

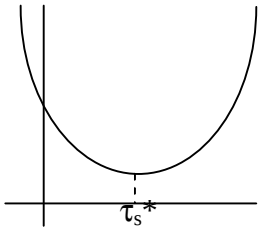
Уравнение (3*) является аналогом уравнений для z и ψ в разностных схемах.

К решению системы (1) применяем метод (2) и посмотрим, как это влияет на невязку:

$$\|r_{s+1}\|_2 \overset{def}{=} \sqrt{(r_{s+1}, r_{s+1})} \rightarrow \|r_{s+1}\|^2 = (r_{s+1}, r_{s+1}) = ((E-\tau_s A)r_s, (E-\tau_s A)r_s) = \underbrace{(r_s, r_s) + \tau_s^2(Ar_s, Ar_s) - 2\tau_s(Ar_s, r_s)}_{F(\tau_s)}$$

При вычислении x_{s+1}, x_s и r_s уже известны, $F(\tau_s) = \alpha\tau_s^2 + \beta\tau_s + \gamma$, $\tau_s = -\dfrac{\beta}{2\alpha}$

$F(\tau_s) \rightarrow \min\limits_{\tau_s \in R}$, ее решение, это вершина параболы: (2^*) $\boxed{\tau_s = \dfrac{(Ar_s, r_s)}{(Ar_s, Ar_s)}}$



Определение:

Методы класса (2) с параметром (2*) называются методами минимальных невязок.

Параметр τ_s для вычисления x_{s+1} выбирают таким образом, чтобы x_{s+1} обеспечивали минимальную невязку.

Исследование сходимости метода.

Рассмотрим 2 случая: а) общий: $\det A \neq 0$; б) частный: $A=A^T>0$.

а) Общий случай.

пусть A –такая. что для нее существует τ: МПИ сходиться, и при этом $G=E-\tau A$: $\|G\|_2<1$. Пусть r_s – невязка метода МПИ, после вычисления x_s: рассмотрим $\left\| \underbrace{(E-\tau A)}_{из\ МПИ} \underbrace{r_s}_{из\ МН} \right\|_2$

$$\left\|(E-\tau A)r_s\right\|_2^2 = \left((E-\tau A)r_s,(E-\tau A)r_s\right) = (r_s,r_s)+\tau^2(Ar_s,Ar_s)-2\tau(Ar_s,r_s) = F(\tau)$$

$$F(\tau) \geq \underbrace{F(\tau)}_{из\ МН} = (r_{s+1},r_{s+1}) = \|r_{s+1}\|_2^2$$

$$\|r_{s+1}\|_2 \leq \|(E-\tau A)r_s\|_2 \leq \underbrace{\|E-\tau A\|_2}_{G} \cdot \|r_s\|_2 \Rightarrow \boxed{\|r_{s+1}\| \leq \|G\| \cdot \|r_s\|_2 \quad (4)}$$

$$(5)\ \|r_{s+1}\| \leq \|G\|^{s+1}\|r_0\|, где\ G = E-\tau A,\ где\ \tau\ из\ МПИ.$$

Исследование поведения погрешности.

$\|r_{s+1}\|_2$ в методе МН; $Az_{s+1} = r_{s+1}$; $Az_0 = r_0$; $z_{s+1} = A^{-1}r_{s+1}$

$$\|z_{s+1}\|_2 \leq \|A^{-1}\|_2 \cdot \|r_{s+1}\|_2 \leq \|A^{-1}\|_2 \cdot \|G\|_2^{s+1}\|r_0\|_2 \leq \underbrace{\|A\| \cdot \|A^{-1}\|}_{=\mu_{\backslash A}}\|G\|^{s+1}\|z_0\| \Rightarrow \boxed{\|z_{s+1}\| \leq \mu_A \cdot \|G\|^{s+1} \cdot \|z_0\| \ (6)}$$

Теорема:

Пусть дана система (1): $Ax=b$ и $\det A \neq 0$. Пусть для A существует τ: МПИ сходится к решению системы (1) и для переходной матрицы $G=E-\tau A$ $\|G\|_2<1 \Rightarrow$ метод минимальных невязок (2) с параметром (2*) сходиться к решению системы (1) с оценками (5) и (6).

б) Частный случай: $A=A^T>0$.

Теорема:

Если матрица $A=A^T>0$, то метод минимальных невязок (2) с параметром (2*):

$$\frac{x_{s+1}-x_s}{\tau_s} + Ax_s = b,\ \tau_s* = \frac{(Ar_s,r_s)}{(Ar_s,Ar_s)}$$

сходится к решению системы $Ax=b$ с оценками:

$$\|r_{s+1}\|_2 \leq \left(\frac{\mu_A-1}{\mu_A+1}\right)^{s+1} \cdot \|r_0\|_2 \quad (5*) \qquad \|z_{s+1}\|_2 \leq \mu_A\left(\frac{\mu_A-1}{\mu_A+1}\right)^{s+1} \cdot \|z_0\|_2 \quad (6*)$$

Доказательство: Для $A=A^T>0$ существует $\tau*$ из МПИ, при котором МПИ сходится и $\|G\|_2<1$, в качестве такого τ, можно брать $\tau* = \frac{2}{\lambda_1+\lambda_n}$, *тогда* $\underbrace{\|G\|_2}_{из\ МПИ} = \frac{\mu_A-1}{\mu_A+1} < 1$

Воспользуемся этими особенностями и оценками (5), (6) $\Rightarrow$ получим (5*) и (6*).

Вывод: для симметричных положительно определенных матриц $A=A^T>0$ мы построили метод, для которого не нужно знать собственные числа матрицы, и который сходиться к точному решению не хуже МПИ с оптимальным параметром.

7.11 Эффективные оценки сходимости (для метода МПИ).

Есть разные приемы изучения сходимости численных методов линейной алгебры. До сих пор при изучении МН и МПИ мы использовали нормы и собственные числа – т.е. инструменты самой линейной алгебры, вместе с тем можно использовать более общие инструменты из функционального анализа – это:

- теоремы об операторе сжатия
- теоремы о неподвижной точке
- теоремы об МПИ в общем случае

На их основе доказывается важные теоремы для МПИ в случае линейных систем:

Теорема 1:

Необходимым и достаточным условием сходимости МПИ к решению системы Ax=b является условие: $\boxed{p(G)<1,\ G=E-\tau A}$

Теорема 2:

Пусть ||G||≤q<1, тогда МПИ сходиться к решению системы Ax=b с любого начального приближения x_0 с эффективными оценками:

$$\|x_{s+1}-x^*\|\le\frac{q}{1-q}\|x_{s+1}-x_s\|\ (7)-\begin{matrix}\textit{априорная}\\ \textit{оценка}\end{matrix}\qquad \|x_{s+1}-x^*\|\le\frac{q^{s+1}}{1-q}\|x_0-x_1\|\ (8)-\begin{matrix}\textit{апостериорная}\\ \textit{оценка}\end{matrix}$$

нормы матрицы должны быть подчинены норме вектора.

Лекция №27. #7.12 Применение метода простой итерации на различных областях.

Рассмотрим разностную схему задачи Дирихле в прямоугольной области:

$A\bar{v}=F\quad(-A)^T=(-A)>0\Rightarrow$ для решения системы $-A\bar{v}=-F$ можно применить МПИ с оптимальным параметром.

Пусть x∈[0,1], y∈[0,1]. Сетка (n,n). Собственные числа матрицы (-A):

$$\lambda_{\min}=\lambda_{11}=\frac{8}{h^2}\sin^2\left(\frac{\pi}{2n}\right)\quad \lambda_{\max}=\lambda_{n-1n-1}=\frac{8}{h^2}\sin^2\left(\frac{\pi(n-1)}{2n}\right)=\frac{8}{h^2}\cos^2\left(\frac{\pi}{2n}\right)$$

число обусловленности: $\mu(-A)=\frac{\lambda_{\max}}{\lambda_{\min}}=ctg^2\left(\frac{\pi}{2n}\right)\approx\frac{1}{2}n^2$. Для МПИ $\tau_{opt}=\frac{2}{\lambda_{\min}+\lambda_{\max}}$.

Метод МПИ в общем виде: $\frac{x_{s+1}+x_s}{\tau}+Ax_s=b\qquad G=E-\tau A-$ переходная матрица.

Для МПИ с оптимальным параметром справедливы оценки:

$$(*)\|z_{s+1}\|_2\le\|G\|_2^{s+1}\|z_0\|_2-\begin{matrix}\textit{из теоремы о}\\ \textit{сходимости}\end{matrix}\qquad (v)\|z_{s+1}\|_2\le\frac{q}{1-q}\|x_{s-1}-x_s\|_2-\begin{matrix}\textit{эффект.}\\ \textit{оценка}\end{matrix}$$

Если $A=A^T>0\Rightarrow\|G\|_2=\frac{\mu_A-1}{\mu_A+1}(**)$. В нашем случае $\|G\|_2=\frac{ctg^2\left(\frac{\pi}{2n}\right)-1}{ctg^2\left(\frac{\pi}{2n}\right)+1}\approx1(***)$

Чем больше n, тем медленнее сходится МПИ с оптимальным параметром.

q – оценка для $\|G\|_2$ – из теоремы 2: $\|G\|_2\le q<1$. Можно использовать в качестве $q=\frac{\mu_A-1}{\mu_A+1}$

Из (v) и (vv) следует:

$$\|z_{s+1}\|_2\le\frac{\mu_A-1}{\mu_A+1}\cdot\frac{\mu_A+1}{(\mu_A+1)-(\mu_A-1)}\|\bar{v}_{s+1}-\bar{v}_s\|_2\Rightarrow\|z_{s+1}\|_2\le\frac{\mu_A-1}{2}\|\bar{v}_{s+1}-\bar{v}_s\|_2\ (vvv)$$

При больших n: $\|\bar{v}_{s+1}-\bar{v}\|\le1/4n^2\|\bar{v}_{s+1}-\bar{v}_s\|_2$

В ситуациях, когда метод практически остановился (на густых сетках), не исключено, что найденное решение $\bar{v}_{s+1}$ все еще сильно отличается от истинного решения системы $\bar{v}$.

Очевидно, чтобы гарантировать $\|\bar{v}_{s+1}-\bar{v}\|\le10^{-6}$ нужно использовать критерий остановки метода $\|\bar{v}_{s+1}-\bar{v}_s\|_2\le10^{-11}$.

Практические выводы: решая линейные системы уравнений итерационным методом, чтобы подойти к решению с заданной точностью, как правило приходиться использовать критерий остановки метода с запасом.

Утверждение:

Если переходная матрица МПИ оценивается $\|G\|_2 \le q \le 1/2$, то из условия $\|x_{s+1} - x_s\| \le \varepsilon$ следует $\|x_{s+1} - x^*\| \le \varepsilon$

Доказательство: $\dfrac{q}{1-q} \le \dfrac{1}{2} \cdot 2 = 1$

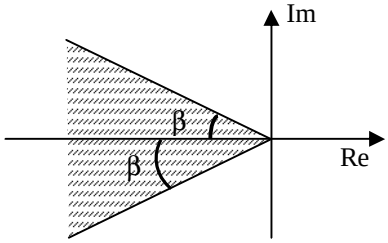
Эти оценки позволяют оценить число итераций для достижения заданной точности. Для построения такой оценки нужно знать $\|G\|_2$. Эта величина не всегда известна и не для всех методов выведены эффективные оценки, поэтому в большинстве случаев действуют наугад, используя наш привычный критерий остановки.

Пример на понимание необходимых и достаточных условий сходимости МПИ.

Заявка: Н.Д.У. МПИ: $\rho(E-\tau A)<1$ (+).

Из (+) следует:

1) Если $A=A^T$, но среди собственных чисел есть и положительные и отрицательные, то МПИ не будет сходиться ни при каком τ.
2) Если A: $\det A \ne 0$ и все ее собственные числа лежат в β - секторе: $\beta \in (0, \pi/2)$, то можно подобрать такое τ, при котором метод будет сходиться.



7.13 задача на отыскание полиномов наименее уклоняющихся от нуля.

Рассмотрим класс функций К: $K = \{a_0 + a_1 x + \ldots + a_{k-1}x^{k-1} + x^k\}$ – полиномы со старшим коэф. 1

Считаем k≥1, задано.

Найти $T_k(x) \in K$, что $\forall P_k(x) \in K$: $\max\limits_{x\in[-1,1]} |T_k(x)| \le \max\limits_{x\in[-1,1]} |P_k(x)|$ (1)

(1) – это классическая задача на отыскание полинома наименее отклоняющегося от нуля.

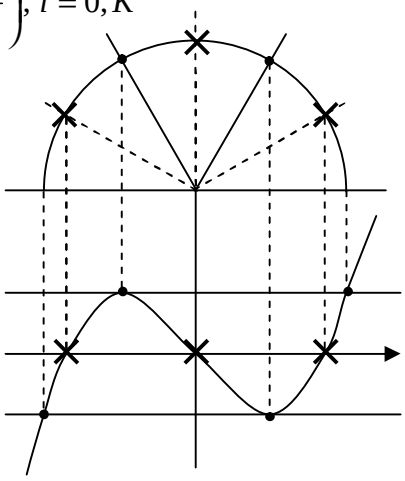
Теорема: Решением задачи (1) является полином Чебышева степени k:

$$(2)\, T_k(x) = (x - x_0)(x - x_1)\ldots(x - x_{k-1}),\ где\ x_s = \cos\left(\frac{\pi}{2k}(1+2s)\right),\ s = 0, k-1$$

причем: $\max\limits_{x\in[-1,1]} |T_k(x)| = \dfrac{1}{2^{K-1}}$ (3) и достигается в точках: $\hat{x}_l = \cos\left(\dfrac{\pi l}{K}\right),\ l = 0, K$

Строим единичную окружность, делим верхнюю дугу на К равных частей и каждую из этих дуг пополам.

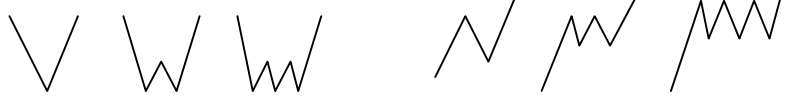
Проекции × дадут корни полинома Чебышева. Проекции • - точки где достигается максимум (3).



Получившийся полином: $T_s(x) = \left(x - \cos\dfrac{\pi}{6}\right)x\left(x - \cos\dfrac{5\pi}{6}\right)$

Из всех $a_0 + a_1 x + a_2 x^2 + x^3$ полином $T_s(x)$ дает самый меньший всплеск и самое сильное падение на [-1,1].

Если К – четное: Если К-нечетное:



К=2 К=4 К=6 К=3 К=5 К=7

Вспомогательная задача на отыскание полиномов наименее отклоняющихся от нуля.

$K=\{1+a_1x+a_2x^2+...+a_Kx^K, a_K\neq 0\}$ *пусть* $[a,b]: 0<a<b.$

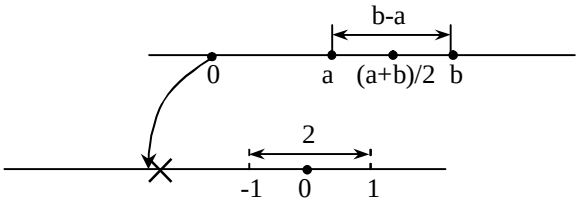
Найдем $T_K(x)\in K$, такой, что $\forall P_K(x)\in K: \max\limits_{x\in[a,b]}|T_k(x)|\leq \max\limits_{x\in[a,b]}|P_k(x)|$ (4)

$T_K(x)$ можно найти из двух условий:

1) $T_K(0)=1$

2) Корни T_K(x) должны быть раскиданы по отрезку [a,b], как корни T_K(x) раскиданы по [-1,1]. Переведем [a,b] в отрезок [-1,1]:

$$\frac{x-\frac{a+b}{2}}{\frac{b-a}{2}};\quad x=0\xrightarrow{b}-\frac{b+a}{b-a}<-1$$



Теорема:

Решением задачи (4) является полином $T_K^{[a,b]}(x): T_K^{[a,b]}(x)\overset{def}{=}\dfrac{T_K\left(\dfrac{x-\frac{a+b}{2}}{\frac{b-a}{2}}\right)}{T_K\left(-\dfrac{b+a}{b-a}\right)}$

Доказательство: по свойству Чебышевского Альтер (??)

Утверждение:

корнем $T_K^{[a,b]}(x)$ являются:

$$\overset{B}{x}_s=\frac{b-a}{2}x_s+\frac{b+a}{2}=\frac{b-a}{2}\cos\frac{\pi}{2K}(1+2s)+\frac{b+a}{2}(5)\quad s=0,k-1;\ x_s\ из\ (2)$$

Доказательство: $T_K(x_s)=0\Rightarrow T_K^{[a,b]}(x)=0,\ если\ \dfrac{x-\frac{a+b}{2}}{\frac{b-a}{2}}$

максимальное по модулю значение $T_K^{[a,b]}$(x) на отрезке [a,b]: $\max\limits_{x\in[a,b]}\left|T_K^{[a,b]}(x)\right|=\dfrac{1}{2^{K-1}}\cdot\dfrac{1}{\left|T_K\left(-\frac{b+a}{b-a}\right)\right|}$.

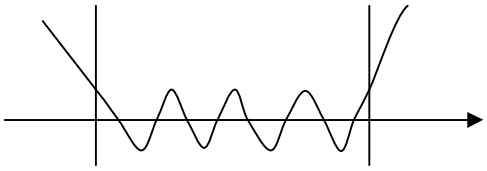
Лекция №28. (01.04.05)

$T_k(x)$ – решение задачи (1) $T_K^{[a,b]}(x)$ – решение задачи (4)

- $T_k(x)$ – полином, который наименее уклоняется от 0 на отрезке [-1,1] в классе К со старшим коэффициентом 1.

- $T_K^{[a,b]}(x)$ – полином Чебышева степени К, который наименее уклоняется от 0, на [a,b] в классе К со свободным коэффициентом 1.

$$T_K^{[a,b]}(x) \overset{def}{=} \frac{T_k\left(\frac{x-\frac{b+a}{2}}{\frac{b-a}{2}}\right)}{T_k\left(-\frac{b+a}{b-a}\right)},$$



- корни $T_K(x)$: x_s, s=0,k-1 и экстремумы достигаются в $\hat{x}_l,\ l=0,K;\ \hat{x}_0=1,\ \hat{x}_k=-1$

- корни $T_K^{[a,b]}(x)$ обозначим как $\hat{x}_s, s=0,K-1$. Экстремальные значения достигаются в $\hat{x}_l, l=0,K$; К-1 локальных экстремумов и $\hat{x}_0=b,\ \hat{x}_K=a$ - в силу граничных условий.

Найдем максимум полинома $T_K^{[a,b]}(x)$, т.е. выясним чему равен $T_K\left(-\frac{b+a}{b-a}\right)$.

Три формы представления классического полинома $T_K(x)$.

$$(I)\ T_K(x)=(x-x_0)(x-x_1)\ldots(x-x_{K-1}),\ x_s - корни$$

$$(II)\ T_K(x)=\frac{1}{2^{K-1}}\cos(k\cdot\arccos x)=x^K+\ldots,\quad x\in[-1,1]$$

$$(III)\ T_K(x)=\frac{\left(x+\sqrt{x^2-1}\right)^K+\left(x-\sqrt{x^2-1}\right)^K}{2^K}=x^K+\ldots,\quad |x|\ge 1.$$

Используя формулу (III) нетрудно показать:

$$\boxed{T_K\left(\frac{b+a}{b-a}\right)=\frac{(-1)^K}{2^K}\frac{1+\rho^{2K}}{\rho^K},\ где\ 0<a<b,\ \rho=\frac{\sqrt{\frac{b}{a}}-1}{\sqrt{\frac{b}{a}}+1}}$$

Доказательство: т.к. 0<a<b, то $\frac{b+a}{b-a}<-1\Rightarrow(III)\Rightarrow\ldots$

Утверждение:

$$\max_{x\in[a,b]}\left|T_K^{[a,b]}(x)\right|=\frac{\max\limits_{x\in[a,b]}\left|T_K(\ldots)\right|\rho^K 2^K}{1+\rho^{2K}}=\frac{2\rho^K}{1+\rho^{2K}}=C$$

Утверждение: если $0<a<b,\ то\ 0<C=\frac{2\rho^K}{1+\rho^{2K}}<1$

Доказательство: $\frac{2\rho^K}{1+\rho^{2K}}<1\Leftrightarrow 2\rho^K<1+\rho^{2K}\Leftrightarrow\rho^{2K}-2\rho^K+1>0\Leftrightarrow\rho^K\ne 1$, т.к.

$$0<a<b,\ то\ \frac{b}{a}>1\ и\ 0<\frac{\sqrt{\frac{b}{a}}-1}{\sqrt{\frac{b}{a}}+1}<1$$

7.14 Метод простой итерации с Чебышевским набором параметров.

Рассмотрим линейную систему Ax=b для матрицы $A=A^T>0$ и $\lambda_i(A)\in[\lambda_1,\lambda_n]$ (1); $\lambda_1>0;\lambda_n>0$

Пусть $\lambda_1 \neq \lambda_n$ – фиксированы.

Рассмотрим класс методов: $\frac{x_{s+1}-x_s}{\tau_s}+Ax_s=b,\ s=0,1,2,\ldots\ (2)$. Пусть К – количество шагов метода.

Изучим, какую погрешность можно получить через К шагов, какое $z_K=x_K-x^*$, если $z_0=x_0-x^*$.

Чем меньше $\frac{\|z_K\|}{\|z_0\|}$, тем лучше.

Будем строить оценку вида: $\|z_K\| \le M(\tau_0,\ldots,\tau_{K-1})\|z_0\|\ (3)$, которая должна быть верна для любых b, x_0, A из (1). И подберем такие параметры $\tau_0,\ldots,\tau_{K-1}$, чтобы число $M(\tau_0^*,\ldots,\tau_{K-1}^*)=\min\limits_{(\tau_0,\ldots,\tau_{K-1})} M(\tau_0,\ldots,\tau_{K-1})$, тогда метод класса (2) с параметрами $\tau_0^*,\ldots,\tau_{K-1}^*$ даст наилучшую гарантию убывания погрешности через К шагов: из (3) $\frac{\|z_K\|}{\|z_0\|} \le M(\tau_0,\ldots,\tau_{K-1})$, поэтому, если минимизировать M - получим наилучшую гарантию.

Рассмотрим конкретную матрицу $A=A^T>0$; для которой $\lambda_i(A)\in[\lambda_1,\lambda_n]$.

$$x_{s+1}=(b-Ax_s)\tau_s+x_s \qquad \underbrace{x_{s+1}-x^*}_{z_{s+1}}=\underbrace{(Ax^*-Ax_s)\tau_s}_{-\tau_s Az_s}+\underbrace{(x_s-x^*)}_{z_s}$$

$$(4)\ z_{s+1}=(E-\tau_s A)z_s=(E-\tau_s A)(E-\tau_{s-1}A)z_{s-1}=\ldots \qquad z_K=\underbrace{(E-\tau_{K-1}A)(E-\tau_{K-2}A)\ldots(E-\tau_0 A)}_{Q_K(\tau_0,\ldots,\tau_{K-1},A)}z_0$$

Это матрица перехода от погрешности z_0 к погрешности z_K.

$$\|z_K\|_2 \le \|Q_K(\tau_0,\ldots,\tau_{K-1},A)\|_2\cdot\|z_0\|_2\ (6)$$

Так как матрица $Q_K=Q_K^T$, то

$$\|Q_K\|_2=\max_{i=1,n}\{|\textit{собственные числа } Q_K|\}=\max_{i=1,n}\{|(1-\tau_{K-1}\cdot\lambda_i(A))(1-\tau_{K-2}\cdot\lambda_i(A))\ldots(1-\tau_0\lambda_i(A))|\}=$$

$$=\boxed{\max_{\lambda\in[\lambda_1,\lambda_n]}\{\underbrace{|(1-\tau_{K-1}\cdot\lambda)(1-\tau_{K-2}\cdot\lambda)\cdot\ldots\cdot(1-\tau_0\lambda)|}_{(8)}\}\ (7)}$$

$(8)=P_K(\lambda)\in K$, т.к. свободный коэффициент =1; корни $P_K(\lambda)$: $\lambda_s=1/\tau_s$, s=0,K-1(9).

Ставим задачу найти такие $\tau_0^*,\ldots,\tau_{K-1}^*$, *чтобы* $\max\limits_{\lambda\in[\lambda_1,\lambda_n]}|P_K(\lambda)|\to\min\ (10)$

нужно выбрать $\tau_0^*,\ldots,\tau_{K-1}^*$ *так, чтобы* $P_K(\lambda)=T_K^{[\lambda_1,\lambda_n]}(\lambda)$, значит:

$$\boxed{\tau_s^*=\frac{1}{\textit{корень}\,T_K^{[\lambda_1,\lambda_n]}(\lambda)};\ \|z_K\|_2\le\max_{\lambda\in[\lambda_1,\lambda_n]}\left|T_K^{[\lambda_1,\lambda_n]}(\lambda)\right|\cdot\|z_0\|_2\ (11)}$$

$$\tau_s=\frac{1}{\frac{\lambda_n+\lambda_1}{2}+\frac{\lambda_n-\lambda_1}{2}\cos\frac{\pi}{2K}(1+2s)},\ s=0,K-1(12)\quad \|z_K\|_2\le\frac{2\rho^K}{1+\rho^{2K}}\|z_0\|,\ \textit{где}\ \rho=\frac{\sqrt{\frac{\lambda_n}{\lambda_1}}-1}{\sqrt{\frac{\lambda_n}{\lambda_1}}+1}=\frac{\sqrt{\mu_A}-1}{\sqrt{\mu_A}+1}(13)$$

Теорема:

Пусть задан класс задач (1): $A=A^T>0$, $\lambda_i(A)\in[\lambda_1,\lambda_n]\Rightarrow$ для любого фиксированного $K\ge1$ существует метод вида (2), который даст наилучшую гарантию убывания погрешности через К шагов, верную для любых b, x_0, A из (1). См. формулы (12) и (13).

Метод (2) с параметрами (12) называется МПИ с чебышевским набором параметров.

Теорема:

При решении задач (1) метод (2) с параметрами (12) можно применять порциями и он будет сходиться к решению системы Ax=b с оценкой $\|z_{KN}\|_2 \le \left(\frac{2\rho^K}{1+\rho^{2K}}\right)^N \|z_0\|_2$ (14), где N – номер порции.

<u>Доказательство:</u> $\frac{2\rho^K}{1+\rho^{2K}} < 1 \Rightarrow \left(\frac{2\rho^K}{1+\rho^{2K}}\right)^N \underset{N\to\infty}{\to} 0$

7.15 Свойства МПИ Чеб.

При К=1 МПИ Чеб совпадает с МПИ с оптимальным набором параметров. МПИ Чеб для К>1 (через N порций расчетов) даст гарантию лучше, чем МПИ с любыми параметрами (следует из выражения (7)). Если в распоряжении 12 итераций, то Чебышевский метод, построенный под К=12 даст гарантию лучше, чем МПИ Чеб (К=3 по 4 раза); (К=4 по 3 раза); (К=2 по 6 раз); (К=6 по 2 раза).

<u>Утверждение:</u> $\frac{2\rho^K}{1+\rho^{2K}} < \left(\frac{\mu_A - 1}{\mu_A + 1}\right)^K$, поэтому МПИ Чеб с К, лучше, чем К итераций МПИ с $\tau_{opt.}$

Лекция №29. (6.04.05)

<u>Свойства МПИ Чеб (К итераций).</u>

<u>1) Сравнение методов МПИ τ_{opt} и МПИ Чеб (К>1) или искусство сравнения методов.</u>

Ax=b $A=A^T>0$ (1)

МПИτ_{opt}: $\frac{x_{s+1}-x_s}{\tau_{opt}} + Ax_s = b, \tau_{opt} = \frac{2}{\lambda_1+\lambda_n}$ (2); $\|z_{s+1}\|_2 \le \left(\frac{\mu_A-1}{\mu_A+1}\right)\cdot\|z_s\|_2$; $\|z_K\|_2 \le \left(\frac{\mu_A-1}{\mu_A+1}\right)^K \|z_0\|_2$ (3)

МПИЧ(К): $\frac{x_{s+1}-x_s}{\tau_s^*} + Ax_s = b,\ x_0 \in R^n$ (4) $\tau_s^*, s = 0, K-1;$ $\|z_K\|_2 \le \frac{2\rho^K}{1+\rho^{2K}}\|z_0\|_2$, $\rho = \frac{\sqrt{\mu_A}-1}{\sqrt{\mu_A}+1}$ (5)

т.к. $\frac{2\rho^K}{1+\rho^{2K}} < \left(\frac{\mu_A-1}{\mu_A+1}\right)^K$, при К>1, (6) то МПИ Чеб(К) даст более сильную гарантию убывания погрешности через К шагов, чем МПИ$\tau_{opt.}$

Аналогично: $\tau_{opt}: \|z_{KN}\|_2 \le \left(\frac{\mu_A-1}{\mu_A+1}\right)^{KN} \|z_0\|_2$ (7); $Чеб(K): \|z_{KN}\|_2 \le \left(\frac{2\rho^K}{1+\rho^{2K}}\right)^N \|z_0\|_2$ (8)

МПИ Чеб(К) при К>1 дает гарантию более быстрой сходимости, чем МПИτ_{opt}

$x<100;\ \hat{x}<1000 \Big| \to$ следует ли отсюда, что $x<\hat{x}$? – нет: 5<100; 1<1000.

Мы работаем с не улучшаемыми оценками! Почему мы сравниваем методы с помощью оценок (3) и (5):

1) Мы говорим о гарантиях сходимости;
2) Чубышевский метод на самом деле лучше, потому, что оценки (3) и (5) основаны на подчиненных нормах и каждая для своего метода является наилучшей, т.е. наиболее точной.

Докажем, что оценка (5) является не улучшаемой оценкой для МПИ Чеб:

Для метода МПИ Чеб(К) была оценка: $z_K = \underbrace{\left(E-\tau_{K-1}^* A\right)\left(E-\tau_{K-2}^* A\right)\ldots\left(E-\tau_0^* A\right)}_{Q_K\left(\tau_0^*,\tau_1^*,\ldots,\tau_K^*,A\right)} z_0$

$$\Rightarrow \|z_K\|_2 \le \underbrace{\left(\max\ из\ модуля\ собств.\ чисел\, Q_K\right)}_{=\|Q_K(\ldots)\|_2}\|z_0\|_2$$

$\frac{\|z_K\|_2}{\|z_0\|_2} \le \max_{i=1,n}\left\{\left|\left(1-\tau^*_{K-1}\lambda_i\right)\ldots\left(1-\tau^*_0\lambda_i\right)\right|\right\} \le (**)$ эту оценку нельзя улучшить, т.к. она основана на подчиненных нормах. $(**) \le \max_{\lambda\in[\lambda_1,\lambda_n]}\left|\underbrace{\left(1-\tau^*_{K-1}\lambda\right)\ldots\left(1-\tau^*_0\lambda\right)}_{(*)}\right| = \max_{\lambda\in[\lambda_1,\lambda_n]}\left|T_K^{[\lambda_1,\lambda_n]}(\lambda)\right| = \frac{2\rho^K}{1+\rho^{2K}}, \rho = \frac{\sqrt{\mu_A}-1}{\sqrt{\mu_A}+1}$

τ* подбирается именно так, чтобы (*) совпадает с полином Чебышева. Но $\max_{\lambda\in[\lambda_1,\lambda_n]}\left|T_K^{[\lambda_1,\lambda_n]}(\lambda)\right|$ достигается в точках $\hat{\lambda}_l, l=0,K$, в том числе в точках λ_1 и λ_n, поэтому вместо знака в (**) стоит знак =.

<u>Вывод:</u> константу оценки (5) улучшить нельзя!

<u>Пояснение:</u> оценка (5) наилучшая с следующем смысле:

$$\forall\varepsilon>0 \Rightarrow \forall A \Rightarrow \exists b, \exists x_0 : \|z_K\| > \frac{2\rho^K}{1+\rho^{2K}} - \varepsilon; \quad \forall A: A = A^T \ \ и \ \ 0 < \lambda_1 \le \ldots \le \lambda_n$$

искусство сравнения методов состоит в том, что нужно сравнивать не улучшаемые оценки.

МПИ Чеб(К) дает оптимальную оценку убывания погрешности за К шагов для <u>класса</u> матриц, а не для <u>конкретной</u> матрицы.

Пример: $\underset{n\times n}{A = A^T} > 0$ – два собственных числа λ_1 (кратн. n_1) и λ_2 (кратн. n_2): $n_1+n_2=n$

например: $0 < \underset{крат.1}{\lambda_1} < \lambda_2 = \underset{крат.\ n-1}{\lambda_3 = \ldots = \lambda_n}$, т.к. матрица симметрична, полный базис у нее есть. Для решения Ax=b можно использовать метод МПИ Чеб(К=2),

$$\frac{x_{s+1}-x_s}{\tau^*_s} + Ax_s = b \ (9) \quad \tau^*_0 = \frac{1}{\frac{\lambda_2+\lambda_1}{2} + \frac{\lambda_2-\lambda_1}{2}\cos\frac{\pi}{4}}; \tau^*_1 = \frac{1}{\frac{\lambda_2+\lambda_1}{2} + \frac{\lambda_2-\lambda_1}{2}\cos\frac{3\pi}{4}}; \ \tau^*_{0,1} = \frac{1}{\frac{\lambda_2+\lambda_1}{2} + \frac{\lambda_2-\lambda_1}{2}\frac{\sqrt{2}}{2}} (9)$$

$$\|z_2\| \le \frac{2\rho^2}{1+\rho^4}\|z_0\|_2, \rho = \frac{\sqrt{\frac{\lambda_2}{\lambda_1}}-1}{\sqrt{\frac{\lambda_2}{\lambda_1}}+1} (10) \text{ рассмотрим метод } \frac{x_{s+1}-x_s}{\tau_s} + Ax_s = b, \tau_0 = \frac{1}{\lambda_1}, \tau_1 = \frac{1}{\lambda_2} (11)$$

$$z_2 = (E-\tau_0 A)(E-\tau_1 A)z_0, \text{ если } \|z_0\| \ne 0, \ то \ \frac{\|z_2\|_2}{\|z_0\|_2} \le \max\left(\underbrace{модулей\ с.ч.\ матрицы\ Q_2}_{(+)}\right)$$

λ_1,λ_2 – собственные числа A.

$$\begin{array}{l|l} кр.1 & (1-\lambda_1\tau_0)(1-\lambda_1\tau_1) - собственные\ числа\ Q_2(\ldots) \\ кр.n-1 & (1-\lambda_2\tau_0)(1-\lambda_2\tau_1) - собственные\ числа\ Q_2(\ldots) \end{array}$$

$$0 = \left(1-\frac{\lambda_1}{\lambda_1}\right)\left(1-\frac{\lambda_1}{\lambda_2}\right) - собственные\ ч.\ Q_2; \quad 0 = \left(1-\frac{\lambda_2}{\lambda_1}\right)\left(1-\frac{\lambda_2}{\lambda_2}\right) - собственные\ ч.\ Q_2$$

⇒ у матрицы Q_2 собственные числа 0 кратности n. ⇒(+)=0 ⇒ метод (11) даст точное решение системы Ax=b за 2 итерации метод (11) не является Чебышевским, но для конкретной матрицы лучше.

<u>Анализ поведения погрешности Чебышевского метода.</u>

Ax=b $A=A^T>0$ (12) $\frac{x_{s+1}-x_s}{\tau^*_s} + Ax_s = b, \ x_0 \in R^n, \tau^*_s, \ s = 0, K-1$

Из x_s получим x_{s+1}, x=0,…,K-1. Мы изучаем поведение погрешности в рамках порций расчетов из К шагов.

$$z_{s+1} = \left(E-\tau^*_s A\right)z_s \ (14) \qquad \|z_{s+1}\|_2 \le \max\left(мод.собст.чис.E-\tau^*_s A\right)\|z_s\|_2 \ (15)$$

$$=\max_{i=1,n}\left\{\left|1-\tau^*_s\lambda_i\right|\right\}$$

$$\max\left\{\left|1-\tau_s^*\lambda_1\right|,\left|1-\tau_s^*\lambda_n\right|\right\}=F\left(\tau_s^*\right)(++)\ из\ \#7,8-7,9$$

Оценка (15) не улучшаемая, т.к. основана на подчиненной норме. В зависимости от $\tau^*_s\ (++)\begin{bmatrix}>1\\=1\\<1\end{bmatrix}$

В зависимости от (++) погрешность может возрастать, убывать или не измениться. Выясним, может ли погрешность возрастать: $\tau_s^*=\dfrac{1}{\frac{\lambda_n+\lambda_1}{2}+\frac{\lambda_n-\lambda_1}{2}\cos\frac{\pi}{2K}(1+2s)}$. Если

$$K>>1\quad \cos\frac{\pi}{2K}\approx 1,\ \cos\frac{\pi\left(2K-1\right)}{2K}\approx -1$$

Для τ_s^*, $s=0,1,2,\ldots$ $\tau_s^*\approx\dfrac{1}{\lambda_n}$ – в начале списка параметров.

Для τ_s^*, $s=0,1,2,\ldots$ $\tau_s^*\approx\dfrac{1}{\lambda_1}$ – в конце списка параметров.

В начале списка: $F\left(\tau_s^*\right)=\max\left\{\approx\left|1-\lambda_1\dfrac{1}{\lambda_n}\right|,\approx\left|1-\lambda_n\dfrac{1}{\lambda_n}\right|\right\}\Rightarrow F\left(\tau_s^*\right)<1$

В конце списка: $F\left(\tau_s^*\right)=\max\left\{\approx\left|1-\dfrac{\lambda_1}{\lambda_1}\right|,\approx\left|1-\dfrac{\lambda_0}{\lambda_1}\right|\right\}\approx F\left(\tau_s^*\right)\approx\mu_A-1$

<u>Утверждение:</u> Если $A=A^T>0$ и $\mu_A>2$, то в методе МПИ Чеб(К) при К>>1 на завершающих итерациях порций расчетов (при $s\approx K-1, K-2, K-3,\ldots$) погрешность может возрастать:

$\left\|z_{s+1}\right\|_2\le F\left(\tau_s^*\right)\left\|z_s\right\|_2$, (18) *где* $F\left(\tau_s^*\right)\approx\mu_A-1$

Д/з: показать, что на последних итерациях порции МПИ Чеб (К), при К>>1, для решения разностной схемы задачи Дирихле, области $[0,1]\times[0,1]$ на густой сетке n=m>>1, за одну итерацию погрешность может возрасти в $n^2/4$.

Пример: (1000×1000) матрица $10^6\times10^6\Rightarrow$ за одну итерацию погрешность может возрасти в $10^6/4$ раз!

Машина может выйти на переполнение разрядной сетки до того, как будет найден x_K, чтобы этого не случилось, используют специальное упорядочивание параметров, так, чтобы итерации на которых погрешность может возрастать, чередовались с итерациями, на которых она убывает [Бахвалов].

Лекция №30. #7.16 Итерационные методы линейной алгебры, основанные на оценках границ спектра матриц.

Во многих прикладных задачах собственные числа матрицы А неизвестны, но с помощью теоремы Гершгорина или численных методов (степенной метод) удается получить оценки границ спектра.

<u>Чебышевский метод, основанный на оценках границ спекта.</u>

МПИ Чеб(К)*: Ax=b (1); $A=A^T>0$: $\lambda^*_1\le\lambda_2\le\ldots\le\lambda^*_n$ (2) – нижняя и верхняя оценка границ спектра.

$$\frac{x_{s+1}-x_s}{\tau_s^{**}}+Ax_s=b,\ x_0\in R^n\ (3)\quad \tau_s^{**}=\frac{1}{\frac{\lambda_1^*+\lambda_n^*}{2}+\frac{\lambda_n^*-\lambda_1^*}{2}\cdot\cos\frac{\pi}{2K}(1+2s)}\quad s=0,K-1$$

Исследуем сходимость метода (3) к решению задачи (1): $z_K=\underbrace{\left(E-\tau_{K-1}^{**}A\right)\cdot\ldots\cdot\left(E-\tau_0^{**}A\right)}_{Q_K(\ldots)}z_0$

Если $\left\|z_0\right\|_2\ne0$, *то*: $\dfrac{\left\|z_K\right\|_2}{\left\|z_0\right\|_2}\le\max\left\{мод.\ соб.\ чисел\ Q_K(\ldots)\right\}=\max\limits_{i=1,n}\left\{\left|\left(1-\tau_{K-1}^{**}\lambda_i\right)\ldots\left(1-\tau_0^{**}\lambda_i\right)\right|\right\}\le$

$$\le \max_{\lambda\in[\lambda_1,\lambda_n]} \left|\left(1-\tau_{K-1}^{**}\lambda\right)\cdot\ldots\cdot\left(1-\tau_0^{**}\lambda\right)\right| \le \max_{\lambda\in[\lambda_1^*,\lambda_n^*]}\left|P_K(\lambda)\right| = \begin{pmatrix}\tau_s^{**}, s=0,K-1 \text{ были подобраны} \\ \text{так, чтобы } P_K(\lambda)=T\ldots\end{pmatrix} =$$

$$\text{при фикс. } \tau_s^{**}, s=0,K-1;\ P_K(\lambda)$$

$$= \max_{\lambda\in[\lambda_1^*,\lambda_n^*]}\left|T_K^{[\lambda_1^*,\lambda_n^*]}\right| = \frac{2(\rho^*)^K}{1+(\rho^*)^{2K}} < 1, \text{ где } \rho^* \overset{def}{=} \frac{\sqrt{\frac{\lambda_n^*}{\lambda_1^*}}-1}{\sqrt{\frac{\lambda_n^*}{\lambda_1^*}}+1}, \text{ причем } 0\le\rho^*<1$$

таким образом: $\dfrac{\|z_K\|_2}{\|z_0\|_2} \le \dfrac{2(\rho^*)^K}{1+(\rho^*)^{2K}} < 1$ (4)

Утверждение:

МПИ Чеб(К)* на основе оценок границ спектра А сходиться к решению задачи Ax=b при любом $x_0\in R^n$ с оценкой (4), из которой следует (5): $\|z_{KN}\|_2 = \left(\dfrac{2(\rho^*)^K}{1+(\rho^*)^{2K}}\right)^N \cdot \|z_0\|_2 \underset{N\to\infty}{\to} 0$

Справка по Чебышевскому методу:

1. Метод является оптимальным для класса матриц, а не для конкретной матрицы.
2. Для каждой конкретной матрицы оценка вида: $\|z_K\|_2 = \dfrac{2\rho^K}{1+\rho^{2K}}\cdot\|z_0\|_2$ является не улучшаемой.

Если метод построен на основе оценок границ спектра, от оценка: $\|z_K\|_2 \le \dfrac{2(\rho^*)^K}{1+(\rho^*)^{2K}}\|z_0\|_2$ не будет не улучшаемой для конкретной матрицы, но она гарантирует сходимость.

Пример: $A = \begin{pmatrix} 9 & 5 & 2 \\ 5 & 13 & 3 \\ 2 & 3 & 7 \end{pmatrix}$ *Круги Гершгорина*: 1) $|z-9|\le 7$ 2) $|z-13|\le 8$ 3) $|z-7|\le 5$ $z\in G$

$A = A^T > 0$

т.к. матрица симметрична ⇒ все ее собственные числа действительны.
{9±7,13±8,7±5} max=21, min=2 ⇒ $\lambda_1^*=2$, $\lambda_3^*=21$.

#7.12. Продолжение. Ликвидация долгов.

Анализ сходимости МПИ в случае несимметричных матриц, а также применения МПИ к решению разностной схемы задачи Дирихле в двумерной области с нелинейной границей.

(1) $Ax=b$, $\det A\ne 0$; (2) $\dfrac{x_{s+1}-x_s}{\tau} + Ax_s = b$

МПИτ: необходимым и достаточным условием сходимости (2) к (1): $\boxed{\rho(E-\tau A)<1}$ (3)

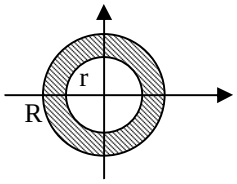
Приемы:

1) если у A; λ_i – собственные числа, i=1,n ⇒ $\tau\lambda_i$ – собственные числа τA, i=1,n.

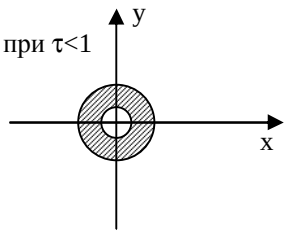
Определение:

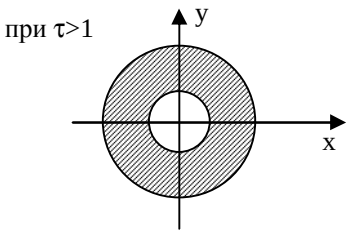
1-$\tau\lambda_i$ – собственные числа E-τA, i=1,n; $\rho(E-\tau A) \overset{def}{=} \max\limits_{i=1,n}(\text{мод. с.ч. } E-\tau A)$

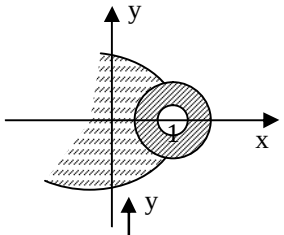
если $(r=\min|\lambda_i|; R=\max|\lambda_i|)$ ⇒ собственные числа A находятся в кольце.



Собственные числа τA в кольце |τ|r, |τ|R.







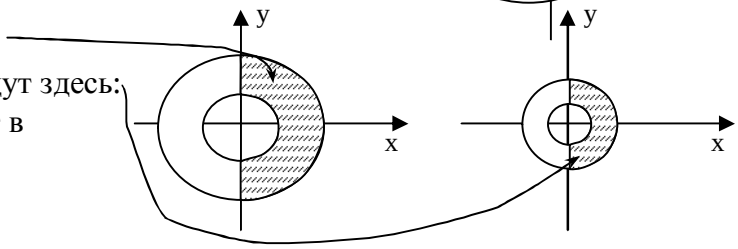
Собственные числа (E-τA). Чтобы проверить условие сходимости (3) нужно чтобы собственные числа E-τA попали в единичный круг (не попадая на его границу).

Если собственные числа матрицы A здесь:

то при τ>0, |τ|<1 собственные числа τA будут здесь:

⇒ собственные числа матрицы E-τA будут в полукольце:

⇒ есть шансы, что собственные числа E-τA попадут в единичные круг.



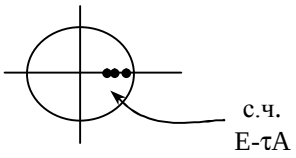
Утверждение:

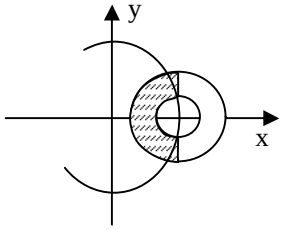
Если все собственные числа A положительны, по существует такое τ, что МПИ сходиться.

Доказательство: $\lambda_i>0$, i=1,n; $\tau\lambda_i>0$; $1-\tau\lambda_i<1$;

При τ>0 достаточно малом получим:

⇒ сходимость есть.





Утверждение 2:

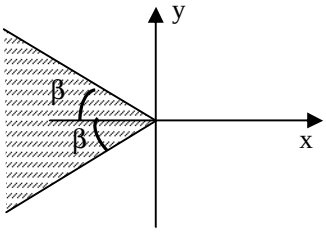
Если A имеет собственные числа λ_i, λ_j такие, что $(\mathrm{Re}\,\lambda_i)(\mathrm{Re}\,\lambda_j)<0$, то МПИ не сходится ни при каком τ.

Утверждение 3:

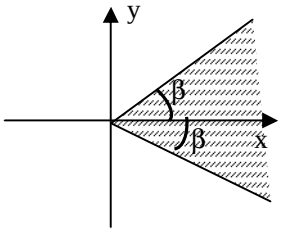
Если матрица A имеет собственные числа ic, -ic, то МПИ не сходится ни при каком τ.

Утверждение, полезное для изучения задачи Дирихле в нелинейной области.

отрицательный β-сектор

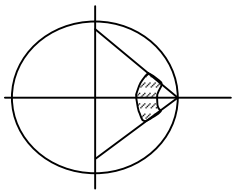


положительный β-сектор.



Утверждение:

Пусть A: detA≠0 такова, что все ее собственные числа лежат в отрицательном β секторе, где β∈(0,π/2)⇒ существует такое τ<0, достаточно малое по модулю, что МПИ будет сходиться.



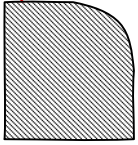
Доказательство: собственные числа будут лежать в пересечении сектора и кольца. Если τ по модулю достаточно мало, то все собственные числа будут внутри единичного круга. ⇒ есть сходимость.

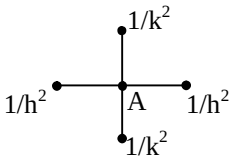
Обоснование применения МПИ к решению задачи Дирихле в двумерной области с нелинейной границей.

$\ddot{A}\bar{v} = F$, *см. рис* #6

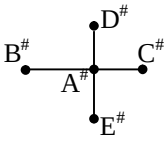
$\ddot{A}$ – *не симметрична* :

$\ddot{A} \neq \ddot{A}^T$





p1

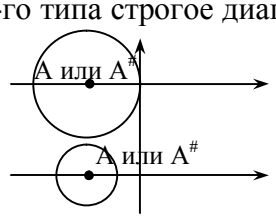


p2

Разностная схема основана на шаблонах: (p1), (p2).

A<0, $A^{\#}$<0 $\left|A^{\#}\right| = \left|B^{\#}\right| + \left|C^{\#}\right| + \left|D^{\#}\right| + \left|E^{\#}\right|$ $A = 2\left(\frac{1}{h^2} + \frac{1}{k^2}\right)$

Поэтому в строках $A_{:,}$, соответствие внутренним узлам 1 типа – нестрогое диагональное преобладание. А в строках, соответствующих внутренним узлам 2-го типа строгое диагональное преобладание.



Теорема Гершгорина:

Круги Гершгорина, соответствующие узлам 1-го типа:
Круги Гершгорина, соответствующие узлам 2-го типа:
Собственные числа лежат внутри или на границе этих кругов, собственных чисел конечное число, 0-го числа нет ⇒ существует отрицательный β-сектор (β∈(0, π/2)) включающий все собственные числа матрицы $A_{:,}$

Теорема:

Разностную схему задачи Дирихле в двумерной области с нелинейной границей: $A_{:,}\bar{v} = F$ можно решить с помощью МПИ, где τ<0 и достаточно малый по модулю.

Д/з: к #7.16 построить МПИ для симметричной положительно определенной матрицы $A=A^T>0$, если известна оценка границы спектра.

#7.17 Теоремы о сходимости метода Зейделя и метода Верхней Релаксации.

Зейдель: Ax=b; A=L+D+R; $(D+L)(x_{s+1}-x_s)+Ax_s=b$; $x_{s+1}-x_s+(D+L)^{-1}Ax_s=(D+L)^{-1}b$.

Теорема:

Метод Зейделя – есть МПИ, примененная к решению системы: $(D+L)^{-1}Ax=(D+L)^{-1}b$.

$\rho\left(E-(D+L)^{-1}A\right)<1$

Лекция №31.

Вместо решения системы Ax=b (1) detA≠0 можно рассматривать и решать систему CAx=Cb (2). Можно строить любой метод для системы (2). За счет удачного выбора матрицы C (detC≠0) можно значительно повысить скорость сходимости.

Как показывают две приведенные выше теоремы, метод Зейделя и метод Верхней Релаксации являются частными случаями этого подхода. Эти теоремы являются самыми главными теоремами о сходимости методов Зейделя и Верхней Релаксации.

Теоремы о сходимости Верхней Релаксации для $A=A^T$, сформированные в декабре прошлого года, являются частными случаями применения этих 2-х теорем.

На основе этих теорем можно получить априорные и апостериорные (эффективные) оценки сходимости методов Зейделя и Верхней Релаксации, аналогичные тем, которые были в #7.12. (Демидович 1950 «ЧМ», Вержбицкий)

#7.18 Свойства сопряженных направлений.

$$(1)\begin{cases} Ax = b,\ x \in R^n \\ A = A^T > 0, x^* - \text{точное решение} \end{cases} \qquad (2)\begin{cases} F(x) = (Ax, x) - 2(b, x) \to \min \\ \text{где } A = A^T, n \times n \quad b \in R^n \end{cases}$$

Утверждение:

Каждая из задач (1) и (2) имеют решение и эти решения совпадают. ∃!x*, ∃!x**, x*=x**.

Доказательство: т.к. detA≠0, то x* ∃ и ! при ∀b; ∀h∈R^n, ∀x∈R^n

$$F(x+h) = (A(x+h), x+h) - 2(b, x+h) = (Ax, x) - 2(b, x) + \underbrace{(Ah, x)}_{=(Ax,h)} + (Ax, h) + (Ah, h) - 2(b, h) =$$

$$\boxed{= F(x) + (Ah, h) + 2(Ax - b, h) \quad (3)}$$

берем x=x* - решение (1), берем ∀h∈R^n, h≠0, рассмотрим x=x*+h

$$(4)\ F(x)=F(x^*+h)=F(x^*)+\underbrace{(Ah,h)}_{>0}+2\underbrace{(Ax^*-b,h)}_{0}\qquad F(x)=F(x^*)+\boxed{>0}>F(x^*)\ m.к. h\neq 0, A>0.$$

$m.o: \forall x\neq x^*\ \ F(x)>F(x^*)\Rightarrow$ решение задачи (2) $\exists!$ и совпадает с x*.

<u>Вывод:</u> вместо решения (1) будем решать задачу (2).

<u>Определение:</u>

Векторы h',h"$\in R^n$ такие, что h'≠0, h"≠0 называются <u>сопряженными, относительно</u> матрицы $A=A^T>0$, если (Ah',h")=0.

h' и h" играют в определение одинаковую роль: $0=(Ah',h'')=(h',A^Th'')=(h',Ah'')=(Ah'',h')=0$

<u>Определение:</u>

Система ненулевых векторов (направлений) $h_0, h_1,\ldots,h_{K-1}\in R^n$ называется <u>взаимно сопряженной относительно матрицы</u> $A=A^T>0$, если $\forall i,j$: $i\neq j$; $i,j=0,K-1$ $(Ah_i, h_j)=0$

пусть BC($A=A^T>0$) – обозначает систему взаимно-сопряженных векторов относительно A.

<u>Утверждение:</u>

Если система ненулевых векторов $h_0,\ldots,h_{K-1}$ является BC($A=A^T>0$), то указанные векторы взаимно-независимы.

<u>Доказательство:</u> от противного:

пусть $\exists h_j:\ \ h_j=\sum\limits_{\substack{i=0\\ i\neq j}}^{K-1}\alpha_i h_i\ \ m.к.\ h_j\neq 0\Rightarrow \exists\alpha_l\neq 0,\ l\neq j$

$$\underbrace{(Ah_j,h_j)}_{>0}=\left(\sum_{i=0}^{K-1}\alpha_i(Ah_i),h_j\right)=\sum_{\substack{i=0\\ i\neq j}}^{K-1}\alpha_i\underbrace{(Ah_i,h_j)}_{=0}=0-\text{противоречие.}$$

<u>Следствие:</u>

Система ненулевых векторов $h_0,h_1,\ldots,h_{n-1}\in R^n$ является BC($A=A^T>0$) образует базис в R^n.

Пусть дано $A=A^T>0$ и $h_0,h_1,\ldots,h_{n-1}$ BC($A=A^T>0$) $k\leq n$, рассмотрим линейное многообразие: пусть x_0 – некоторый фиксированный вектор из R^n.

$$L(x_0,h_0,\ldots,h_{n-1})\overset{def}{=}\{x_0+\underbrace{\alpha_0h_0+\alpha_1h_1+\ldots+\alpha_{k-1}h_{k-1}}_{k\ \text{штук степеней свободы}}\},x_0\in R^n,\alpha_i\in R, i=0,K-1$$

$$x=x_0+\sum_{i=0}^{k-1}\alpha_ih_i-\textit{элемент линейного многообразия}$$

$$\textit{рассмотрим } F(x),\textit{при } x\in L(\ldots)\ \ F(x)=F\left(x_0+\sum_{i=0}^{k-1}\alpha_ih_i\right)\overset{(2)}{=}F(x_0)+\left(A\sum_{i=0}^{k-1}\alpha_ih_i,\sum_{i=0}^{k-1}\alpha_ih_i\right)+$$

$$+2\left(Ax_0-b,\sum_{i=0}^{k-1}\alpha_ih_i\right)\textit{исп BC}=\boxed{F(x_0)+\sum_{i=0}^{k-1}\left(\alpha_i^2(Ah_i,h_i)+2(Ax_0-b,h_i)\alpha_i\right)}\ (5)$$

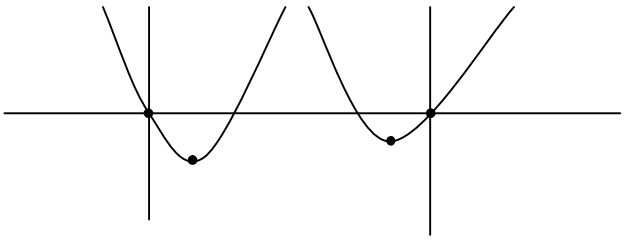
$$F(x)\to\min_{x\in L(\ldots)},x\in R^n$$

Каждое слагаемое в выражении (5) можно минимизировать по отдельности, потому, что каждое зависит только от своего сопряженного направления h_i. Если многообразие L(…) задано, то h_i – фиксированы и каждое слагаемое (5) является параболой относительно α_i. Парабола имеет вид

$\alpha_i^2A+\alpha_i\cdot 2B$

Минимум достигается в вершине:

$$\alpha_i=-\frac{(Ax_0-b,h_i)}{(Ah_i,h_i)}$$



Теорема:

Пусть $A=A^T>0$ и h_i – ВС($A=A^T>0$) и пусть задан $x_0\in R^n$ и линейное многообразие $L(x_0,h_1,\ldots h_{n-1})$ тогда решением задачи: $F(x)\to\min, x\in L(x_0,h_0,\ldots,h_{n-1})$ является такой $x=x_0+\sum_{i=0}^{n-1}x_ih_i$, где при каждом i значение α_i являются решением задачи одномерной минимизации:

$$\alpha_i^2(Ah_i,h_i)+2(Ax_0-b,h_i)\alpha_i\to\min_{\alpha_i\in R}(6)\quad \alpha_i=-\frac{(Ax_0-b,h_i)}{(Ah_i,h_i)}, i=0,k-1(7)$$

Следствие:

Пусть $A=A^T>0$ и $h_0,h_1,\ldots,h_{n-1}$ - ВС($A=A^T>0$) $\Rightarrow$ решением задачи (2) является $x^*\in R^n$:

$$x^*=x_0+\sum_{i=1}^{n-1}\alpha_ih_i, где\ \alpha_i=-\frac{(Ax_0-b,h_i)}{(Ah_i,h_i)}, i=0,n-1$$

Пояснение к доказательству:

$h_0,h_1,\ldots,h_{n-1}$ – базис в $R^n\Rightarrow\forall x\in R^n$ и $\forall x_0\in R^n$

$$x-x_0=\sum_{i=0}^{n-1}\alpha_ih_i\Rightarrow x=x_0+\sum_{i=0}^{n-1}\alpha_ih_i\ и\ поэтому\ L(x_0,h_0,\ldots,h_{n-1})\equiv R^n\ (8)$$

Выводы:

1) Если для $A=A^T>0$ были известны n – штук взаимно-сопряженных направлений, то задачу многомерной минимизации R^n можно было бы свести к решению n задач одномерной минимизации: $\alpha_i^2(Ah_i,h_i)+2(Ax_0-b,h_i)\alpha_i\to\min_{\alpha_i\in R}$
2) Если бы были известны h_i, то решение задач (1) и (2 можно было бы сразу найти по формуле: $\alpha_i=-\frac{(Ax_0-b,h_i)}{(Ah_i,h_i)}(8*)$
3) Если бы для матрицы $A=A^T>0$ h_i, i=0,n-1, x_0 был бы не нужен: $x=\sum_{i=0}^{n-1}\alpha_ih_i,\quad \alpha_i=\frac{(b,h_i)}{(Ah_i,h_i)}$ (9)

7.19 Метод Сопряженных градиентов.

Вместо решения задачи (1) решаем задачу оптимизации (2). Система сопряженных направлений строиться поэтапно в ходе работы метода. В качестве x_0 можно брать $x_0=0$. Чтобы быстрее выйти на точное решение выбирают $x_0\approx x^*$ - точное решение (задачи (1) или задачи (2)).

Лекция №32.

$$(1)\begin{cases}Ax=b,\ x\in R^n\\ A=A^T>0, x^*-точное\ решение\end{cases}\qquad (2)\begin{cases}F(x)=(Ax,x)-2(b,x)\to\min\\ где\ A=A^T, n\times n\quad b\in R^n\end{cases}$$

$x_0\in R^n$ – начальное приближение.

Метод: $x_0\in R^n$; $x_1=x_0+\alpha_0h_0$, $h_0=-r_0=-(Ax_0-b)$.

α_0 – решение задачи минимизации по направлению $F(x_0+\alpha_0h_0)\to\min_{\alpha_0\in R}$ x_0,h_0 – заданы.

$$F(x_0+\alpha_0h_0)=F(x_0)+\alpha_0^2(Ah_0,h_0)+2(Ax_0-b,h_0)\alpha_0\to\min_{\alpha_0\in R}$$

$$\alpha_0=-\frac{(Ax_0-b,h_0)}{(Ah_0,h_0)}=-\frac{(r_0,h_0)}{(Ah_0,h_0)}(4)\qquad x_{s+1}=x_s+\alpha_sh_s, где\ h_s=-r_s+\beta_sh_{s-1},\ r_s\overset{def}{=}Ax_s-b$$

β_s берут так, чтобы направления h_s и h_{s-1} были сопряженными относительно $A=A^T>0$:

$$(Ah_{s-1},h_s)=0=(Ah_{s-1},-r_s+\beta_sh_{s-1})=-(Ah_{s-1},r_s)+\beta_s(Ah_{s-1},h_{s-1})=0\Rightarrow\beta_s=\frac{(Ah_{s-1},r_s)}{(Ah_{s-1},h_{s-1})}(5)$$

α_s берут как решение задачи минимизации по направлению: $F(x_s+\alpha_s h_s)\to\min\limits_{\alpha_s\in R}$, *где* x_s *и* h_s – уже заданы. $F(x_s+\alpha_s h_s)=F(x_s)+\alpha_s^2(Ah_s,h_s)+2(Ax_s-b,h_s)\alpha_s\to\min\limits_{\alpha_s\in R}$

$$\alpha_s=-\frac{(Ax_s-b,h_s)}{(Ah_s,h_s)}=-\frac{(r_s,h_s)}{(Ah_s,h_s)}\quad(6)$$

Теоремы о свойствах метода:

Утверждение: Если x_s такой, что $r_s=0\Rightarrow x_s=x^*$(точное решение).

Теорема:

Если $r_0,\dots,r_s\neq0$, то векторы $h_0,\dots,h_s$ – ВС($A=A^T>0$), а векторы $r_0,\dots,r_s$ взаимно ортогональны: $(r_i,r_j)=0$, $i\neq j$.

Доказательство: смотрите Бахвалов.

Утверждение:

Если $r_0,\dots,r_s\neq0$, то (6) совпадает с (7) из #7.18.

Доказательство: $(Ax_s-b,h_s)=\left(A\left(x_0+\sum\limits_{i=0}^{s-1}\alpha_i h_i\right),h_s\right)-(b,h_s)=(Ax_0-b,h_s)$.

Значит в формуле МСГ: $\boxed{\alpha_s=-\frac{(Ax_0-b,h_s)}{(Ah_s,h_s)}}\quad(7*)$

Вольное утверждение: МСГ генерирует систему ВС($A=A^T>0$). На каждом шаге метода вектор $x_s=x_{s-1}+\alpha_{s-1}h_{s-1}$. На самом деле является вектором $x_s=x_0+\sum\limits_{i=0}^{s-1}\alpha_i h_i$ и обеспечивает min F(x) на линейном многообразии $x\in L(x_0,h_0,\dots,h_{s-1})$. Это очевидно из #7.18.

Утверждение: Не позже, чем на шаге s=n-1, метод даст точное решение задачи $F(x)\to\min$, $x\in R$, т.е. решение системы Ax=b.

МСГ с практической точки зрения:

1) МСГ можно рассматривать, как прямой метод линейной алгебры, т.к. он даст точное решение системы не больше, чем через n шагов.
2) Если n велико, то метод можно рассматривать как итерационный, постепенно приближающийся к точке глобального минимума.
3) При машинном счете вектора $h_0,\dots,h_{s-1}$ в МСГ через некоторое число шагов теряют свойство сопряженности и метод начинает «гулять». Чтобы исправить ситуацию, метод запускают порциями по К шагов, аналогично МПИ Чеб(К). При этом есть теорема о сходимости:

Определение:

$\|x\|_A=\sqrt{(Ax,x)}$, *если* $A=A^T>0$ $\|x\|_A$ – *энергетическая норма.*

Теорема:

Если МСГ применить порциями по К шагов, то погрешность $z_K=x_K-x^*$, где x* - точное решение (1) и (2) в энергетической норме убывает не хуже, чем погрешность Чебышевского метода в евклидовой норме: $\|z_K\|_A\le\frac{2\rho^K}{1+\rho^{2K}}\|z_0\|_A$, где ρ - выражается через μ_A.

Доказательство: в книге Марчука.

Лекция №33. #7.20 Итоги.

Есть разные подходы к решению систем уравнений:

а) можно решать систему;

б) решать задачу оптимизации;

в) метод установления – решать систему ДУ;

$(1)\ Ax = b,\ x^* - решение\quad (2)\dfrac{dx}{dt} = Ax - b,\ x(t) \in R^n(t) \Rightarrow x^*$ – точка равновесия в системе (2).

1. Вместо того, чтобы решать (1), с помощью ЧМ выстраивают траекторию системы (2)
   -Если x* - является устойчивым равновесием системы (2), то траектории x(t) → x*, t→∞ будут сходиться к этому равновесию. Условие применения метода: $\mathrm{Re}(\lambda_i(A)) < 0\ i = 1, n$
2. При численном решении задач математической физики приходиться решать системы линейных уравнений большой размерности с разреженными матрицами. Для решения таких систем, начиная с 60-х, 70-х годов, разрабатывались специальные методы линейной алгебры:
   a. экономичные по числу действий;
   b. вычислительно-устойчивые;
   c. простые в программировании.

   С конкретными методами можно познакомиться в книгах Самарского, Марчука. К этому классу метод относятся: попеременно-треугольный, метод переменных направлений, методы расщепления и др.
   В тех же книгах есть полезные общие теоремы о проверке условий сходимости методов.
3. Полезные теоремы об условии сходимости старых методов: Зейделя, Якоби и простой итерации можно найти в книгах Демидовича, Вержбицкого (2003?)
   В эпоху ручных вычислений итерационные методы были полезны даже для решения систем малой размерности:
   Пример для системы 10×10 метод Гаусса требует примерно 1000 действий, а метод Зейделя или Якоби может дать хорошие приближения решения за меньшее число действий.
4. Разные типы норм: $\|x\|_\infty \quad \|x\|_1 \quad \|x\|_2 \quad \|x\|_A$
   $\|A\|_\infty \quad \|A\|_1 \quad \|A\|_2 \quad \|A\|_A$
   прост.для выч. *Th* *СГ*

   В численных методах линейной алгебры используются разные нормы, причем матричные нормы должны быть подчинены векторным.
   В основном во всех основных теоремах #7 использовалась евклидова норма.
   $\|x\|_\infty$ - обычно используют в программе для вычисления нормы погрешности.
   В книгах есть достаточно много теорем, формулирующих условие сходимости методов в произвольных матричных нормах. Нормы $\|A\|_\infty$ *и* $\|A\|_1$ элементарно вычисляются для любой квадратной матрицы позволяют быстро проверить сходится метод в нормах $\|x\|_\infty$ *или* $\|x\|_1$ или не сходится, т.е. эти нормы нужны для удобства проверки сходимости методов. Если метод сходится в какой-то норме, то он будет сходиться и в любой другой.

   $$\|x\|_\infty = \max_{i=1,n}|x_i| \quad \|x\|_2 = \sqrt{(x,x)} = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$$

   $$\left(\|x\|_\infty\right)^2 n \ge \left(\|x\|_2\right)^2 \ge \left(\|x\|_\infty\right)^2 \Rightarrow \|x\|_\infty \le \|x\|_2 \le \sqrt{n}\,\|x\|_\infty \quad (*)$$

   Эта оценка (*) позволяет пересчитывать эффективные оценки сходимости из одной нормы в другую. Пример: пусть получилось, что
   $\left\|z^{(s)}\right\|_\infty = 0.3\cdot 10^{-6} \Rightarrow \|z\|_2 \le \sqrt{n}\cdot 0.3\cdot 10^{-6}$, *где n – размерность вектора.*

Метод наименьших квадратов. Метод Рунге-Кутта. Метод Бубнова-Галеркина.

#8.1 Метод Наименьших Квадратов.

Есть разные способы приближения функции – заданной формулой или набором точек (интерполяционный полином, сплайн). В этих методах предполагается, что будет построена ли

ния, проходящая через заданные точки.

МНК кардинально отличается от этих подходов, тем, что по набору точек строится линия, которая хорошо приближает набор данных, но не обязана проходить через заданные точки.

Рассмотрим 2 примера: про поплавок и про пенициллиновый бульон. Данные для поплавка можно интерполировать сплайном или полиномом, пенициллиновые данные такими методами обрабатывать нельзя.

И тот, и другой набор данных содержит ошибки эксперимента и некоторую случайность, вызванную наличием неучтенных факторов.

При обработке данных нужно игнорировать случайность и неучтенные факты и строить зависимость y(x) по принципу МНК.

(x_i,y_i), i=1,n – данные (1); x – объясняющая переменная; y – объяснимая переменная.

n – число наблюдений.

$y = a_0 + a_1 x + a_2 x^2 + \ldots + a_k x^k$ – полином степени k (2)

y_i – истинное значение y; $\hat{y}_i$ – оценочное значение y.

$$\hat{y}_i \overset{def}{=} a_0 + a_1 x_i + \ldots + a_k x_i^k \ (3) \qquad \hat{\varepsilon}_i \overset{def}{=} y_i - \hat{y}_i \ - \textit{остаток} \ (4)$$

Принцип наименьших квадратов

Уравнение (2) для набора данных (1) подбирают таким образом, чтобы сумма квадратов остатков была минимальной:

$$S(a_0,\ldots,a_k) \overset{def}{=} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n} \left(y_i - \left(a_0 + a_1 x_i + \ldots + a_k x_i^k\right)\right)^2 \quad (8)$$

$$S(a_0,\ldots,a_k) \to \min_{(a_0,\ldots,a_k)\in R^{k+1}} (6)$$

Функционал S вычисляется для фиксированного набора данных (1) и является квадратной функцией относительно a_j, $j = 0,k$. Нужно построить уравнение (2) для данных (1) по принципу МНК. Необходимым условием минимума в критерии (6) является: $\dfrac{\partial S}{\partial a_j} = 0,\ j = 0,k$ (7)

Теорема:

Для фиксированного набора данных (1), удовлетворяющих условию (*), уравнение вида (2), обеспечивающее минимум суммы квадратов остатков существует и единственно, а его коэффициенты a_j можно найти, решив систему уравнений (7).

(7) – это нормальная система уравнений:

$$\begin{bmatrix} 1 & x_1 & x_1^2 & \mathrm{i} & x_1^k \\ 1 & x_2 & x_2^2 & \mathrm{i} & x_2^k \\ \mathrm{j} & \mathrm{j} & \mathrm{j} & \mathrm{I} & \mathrm{j} \\ 1 & x_n & x_n^2 & \mathrm{i} & x_n^k \\ x_{(0)} & x_{(1)} & x_{(2)} & & x_{(k)} \end{bmatrix} (13^*) - \text{расширенная матрица данных}$$

Условие (*):

(1) должен быть таков, чтобы векторы $\bar{x}_{(0)}, \bar{x}_{(1)}, \ldots, \bar{x}_{(k)}$ (8) были линейно – независимы.

Покажем, что система (7) сводиться к линейной системы уравнений вида:

$$\begin{pmatrix} \left(\bar{x}_{(0)}, \bar{x}_{(0)}\right) & \mathrm{h} & \left(\bar{x}_{(0)}, \bar{x}_{(k)}\right) \\ \mathrm{j} & \mathrm{I} & \mathrm{j} \\ \left(\bar{x}_{(k)}, \bar{x}_{(0)}\right) & \mathrm{i} & \left(\bar{x}_{(k)}, \bar{x}_{(k)}\right) \end{pmatrix} \cdot \begin{pmatrix} a_0 \\ a_1 \\ \mathrm{j} \\ a_k \end{pmatrix} = \begin{pmatrix} \left(\bar{x}_{(0)}, \bar{y}\right) \\ \mathrm{j} \\ \left(\bar{x}_{(k)}, \bar{y}\right) \end{pmatrix} (9) \qquad \bar{y} \overset{def}{=} \left(y_1, y_2, \ldots, y_n\right)$$

Если векторы (8) линейно независимы, то матрица в системе (9) является матрицей Грамма этих векторов, не вырождена и положительно определена, → система (9) имеет единственное решение.

Лекция №34.

Доказательство: (x_i, y_i) i=1,n – известный набор данных (1)

$y = a_0 + a_1 x + ... + a_k x^k$ – искомое уравнение (2).

y_i, $i = 1, n$ истинные значения y. $\hat{y}_i$ (оценочные значения y) : $\hat{y}_i \overset{def}{=} a_0 + a_1 x_i + ... + a_k x_i^k$, $i = 1, n$

По данным (1) строим уравнение (2) так, чтобы: $S(a_0, ..., a_k) \to \min\limits_{(a_0,...,a_k) \in R^{k+1}} (6)$, где

$S(a_0, ..., a_k) \overset{def}{=} \sum\limits_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$ *см.*(5) нужно найти α_j, обеспечивающие минимум S.

Определения:

1) $\underset{\text{градиент вектор}}{S'(a_0, ..., a_k)} = \left[\dfrac{\partial S}{\partial a_0}, ..., \dfrac{\partial S}{\partial a_k}\right]^T$

2) $\underset{\text{матрица 2-х производных}}{S''(a_0, ..., a_k)} = \begin{pmatrix} \dfrac{\partial^2 S}{\partial a_0^2} & \text{h} & \dfrac{\partial^2 S}{\partial a_0 \partial a_k} \\ \text{j} & \text{l} & \text{j} \\ \dfrac{\partial^2 S}{\partial a_k \partial a_0} & \text{i} & \dfrac{\partial^2 S}{\partial a_k^2} \end{pmatrix}$

3) Условие $S'(a_0, ..., a_k) = 0$ является необходимым условием локального экстремума и позволяет найти точки подозрительные на локальный экстремум.

если $S'(a_0^*, ..., a_k^*) = 0$ *и* $S''(a_0^*, ..., a_k^*) > 0$, *то* $(a_0^*, ..., a_k^*)$ – точка локального минимума.

4) Пусть $S(a_0, ..., a_k)$ - квадратичный функционал своих аргументов. Пусть в точке ($a_0^*, ..., a_k^*$) выполняются условия S'=0 и S">0 ⇒ это точка глобального минимума этой функции S.

Доказательство: Для любой точки ($a_0^*, ..., a_k^*$) и для любой точки $(a_0, ..., a_k) \in R^{k+1}$ верно представление:

$$S(a_0, ..., a_k) = S(a_0^*, ..., a_k^*) + \left(S'(a_0^*, ..., a_k^*), (h_0, ..., h_k)^T\right) + \frac{1}{2}\left(S''(a_0^*, ..., a_k^*)(h_0, ..., h_k)^T, (h_0, ..., h_k)^T\right), (10)$$

где $h_j = a_j - a_j^*$, $j = 0, k$

(10) – это разложение функции S в точке ($a_0^*, ..., a_k^*$), других слагаемых в ряду нет, т.к. формула квадратичная. Если $(a_0^*, ..., a_k^*) \neq (a_0, ..., a_k)$, *то* $(h_0, ..., h_k) \neq 0$, $\left(S'' h^T, h^T\right) > 0$

(11) $S(a_0, ..., a_k) = S(a_0^*, ..., a_k^*) + 0 + (>0) > S(a_0^*, ..., a_k^*)$

Из (11) следует, что ($a_0^*, ..., a_k^*$) – точка глобального минимума. Таким образом, при доказательстве теоремы про МНК достаточно найти точку, в которой S'=0 и проверить знак S".

$S(a_0, ..., a_k) \overset{def}{=} \sum\limits_{i=1}^{n} \left(y_i - \left(a_0 + a_1 x_1 + ... + a_k x_i^k\right)\right)^2 \to \min$ S – квадратичная относительно a_j.

$$\frac{\partial S}{\partial a_j} = 2\sum_{i=1}^{n} \left(y_i - \left(a_0 + a_1 x_i + ... + a_k x_i^k\right)\right)(x_i)^j, \ j = 0, k$$

$$\frac{\partial S}{\partial a_j} = 0, \ j = 0, k \ (12) \Rightarrow \sum_{i=1}^{n} \left(a_0 + a_1 x_i + ... + a_k x_i^i\right)(x_i)^j = \sum_{i=1}^{n} y_i (x_i)^j, \ j = 0, k$$

Перепишем уравнение таким образом, чтобы было видно, что это линейная система уравнений, относительно a_j: $(13)\; a_0 \underbrace{\sum_{i=1}^{n}(x_i)^j}_{(\bar{x}_0,\bar{x}_j)} + a_1 \underbrace{\sum_{i=1}^{n}(x_i)^{j+1}}_{(\bar{x}_1,\bar{x}_j)} + \ldots + a_k \underbrace{\sum_{i=1}^{n}(x_i)^{j+k}}_{(\bar{x}_k,\bar{x}_j)} = \underbrace{\sum_{i=1}^{n} y_i x_i}_{(\bar{x}_j,\bar{y})}, \quad j = 0,k$

Для нахождения точки, в которой S'=0б нужно решить систему (13).
Очевидно, что система (13) в матричной форме совпадает с системой (9):

$$\begin{pmatrix} (\bar{x}_{(0)},\bar{x}_{(0)}) & \text{h} & (\bar{x}_{(0)},\bar{x}_{(k)}) \\ \text{j} & \text{l} & \text{j} \\ (\bar{x}_{(k)},\bar{x}_{(0)}) & \text{i} & (\bar{x}_{(k)},\bar{x}_{(k)}) \end{pmatrix} \cdot \begin{pmatrix} a_0 \\ a_1 \\ \text{j} \\ a_k \end{pmatrix} = \begin{pmatrix} (\bar{x}_{(0)},\bar{y}) \\ \text{j} \\ (\bar{x}_{(k)},\bar{y}) \end{pmatrix} \quad (9) \qquad \bar{y} \overset{def}{=} (y_1, y_2, \ldots, y_n)$$

Матрица системы (9) является матрицей Грамма системы векторов $(\bar{x}_{(0)},\ldots,\bar{x}_{(k)})$ (*). В силу линейной независимости этих векторов (*), матрица Грамма не вырождена и положительна определена. Gr>0 и detGr≠0 ⇒ система (9) имеет единственное решение ⇒ есть точка подозрительная на локальный экстремум. Точка $(a_0^*,\ldots,a_k^*)$ - существует, единственна и подозрительна на локальный экстремум. Найдем $S''(a_0,\ldots,a_k)$ и проверим положительную определенность матрицы: $\dfrac{\partial^2 S}{\partial a_j \partial a_l} = -2\sum_{i=1}^{n}(x_i)^j \underbrace{(x_i)^l(-1)}_{\text{результат дифф.}} = 2\sum_{i=1}^{n}(x_i)^{j+l} = 2(x_{(j)},x_{(l)})$ это элемент матрицы S" в строке j и столбце l ⇒ $S''(a_0^*,\ldots,a_k^*) = S'' = 2Gr > 0 \Rightarrow (a_0^*,\ldots,a_k^*)$ – точка глобального минимума. Теорема доказана.

МНК на практике

1. Чтобы построить по набору данных (x_i,y_i) i=1,n (1) зависимость вида $y = a_0 + a_1 x_i + \ldots + a_k x_i^k$ (2) нужно построить расширенную матрицу данных:

$$\bar{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \text{i} & x_1^k \\ 1 & x_2 & x_2^2 & \text{i} & x_2^k \\ \text{j} & \text{j} & \text{j} & \text{l} & \text{j} \\ 1 & x_n & x_n^2 & \text{i} & x_n^k \end{bmatrix}$$

и убедиться в том, что ее столбцы линейно независимы, а затем решить систему (9);

2. Если в наборе данных при ∀i≠j $x_i \neq x_j$ – все различны, то столбцы матрицы (9) линейно независимы!

3. Задачу можно решить по определению – т.е. выписать формулу S и решить систему $\dfrac{\partial S}{\partial a_j} = 0, \;\; j = 0,k$

Критерий качества построенного решения.

Пусть дан набор данных (x_i, y_i) i=1,n и построен МНК-полином $y = a_0 + a_1 x + \ldots + a_k x^k$. Чтобы определить насколько хорошо этот полином подходит к данным используют характеристики:

1. Доля объясненной дисперсии.

$$\boxed{R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \hat{y}_{средн})^2}{\sum_{i=1}^{n}(y_i - y_{сред})^2}}$$, (коэффициент детерминации). $y_{сред} = \dfrac{\sum_{i=1}^{n} y_i}{n}$, $\hat{y}_{сред} = \dfrac{\sum_{i=1}^{n}\hat{y}_i}{n}$

Известно: $0 \le R^2 \le 1$. Чем ближе R^2 к 1, тем лучше (2) соответствует (1).

2. Стандартная ошибка оценки:

$\boxed{s=\sqrt{\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)}{n-(k+1)}}}$, n – число пар данных, k – количество искомых переменных. Чем меньше S тем лучше.

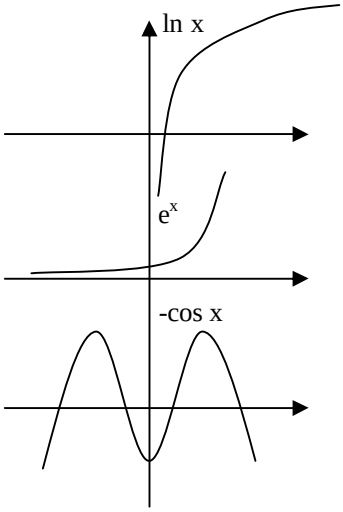
3. $\hat{\varepsilon}_i = y_i - \hat{y}_i$ – остаток. На графиках остатков в осях y-$\hat{\varepsilon}$ или $\hat{y}-\hat{\varepsilon}$ должны отсутствовать закономерности. Если присутствует какая-либо закономерность, значит существует какая-либо зависимость y(x), не учтенная уравнением (2).
4. Используются статистические критерии качества (значимость, автокорреляцию и другие факторы).

Выбор k для заданного n.

1. Если k=n-1, то МНК – полином совпадает с интерполирующим полиномом (доказать самим из здравого смысла).
2. Если k достаточно большое, то даже при выполнении условия $\forall i\neq j\ x_i\neq x_j$ окажутся почти линейно зависимы, матрица Грамма будет плохо обусловленной. Система (9) будет решена с погрешностью, поэтому большие k не используют.
3. По мнению инженеров, большие k не нужны, из здравого смысла.

Пример: Степень полинома k подбирают до тех пор, пока не получат случайный график остатков.

Лекция №35. #8.1 МНК-приближение с помощью обобщенных полиномов степени k.



Рассмотрим ситуацию, когда набор данных не похож на полином:

$\hat{y} = a_0\cdot\boxed{1}+a_1\boxed{x^1}+\ldots+a_k\boxed{x^k}\ (*)$ для данных задачи $S(a_0,\ldots,a_k)\to\min$ определяются тем, что S квадратично зависит от $a_j,\ j=0,k$ и решение задачи сводиться к решению линейной системы уравнений $\frac{\partial S}{\partial a_j}=0$ и проверки условия S”>0.

Эти свойства не зависят от того, какие функции от x используются в выражении (*), т.е. вместо функций: 1, x, x^2, …, x^k можно использовать: $\varphi_0(x)$, $\varphi_1(x)$,…,$\varphi_k(x)$. В выражении (*) $\varphi_0(x)=1$, $\varphi_1(x)=x$,…, $\varphi_k(x)=x^k$.

Определение:

Обобщенным полиномом степени k относительно x называют выражение: $\hat{y}=a_0\varphi_0(x)+a_1\varphi_1(x)+\ldots+a_k\varphi_k(x)$ (14)

$x_i, y_i, i=1,n$ – данные задачи (1); y_i, i=1,n – истинные значения y; $\hat{y}_i, i=1,n$ – оценочные значения.

$\hat{y}_i=\sum_{j=0}^{k}a_j\varphi_j(x_i)$ (15); $\hat{\varepsilon}_i \overset{def}{=} y_i-\hat{y}_i$ – *остатки*; $S(a_0,\ldots,a_k)\to\min_{a_0,\ldots,a_k\in R^{k+1}}$ (16)

$$S(a_0,\ldots,a_k)\overset{def}{=}\sum_{i=1}^{n}\left(\hat{\varepsilon}_i\right)^2=\sum_{i=1}^{n}\left(y_i-\sum_{j=0}^{k}a_j\cdot\varphi_j(x_i)\right)^2 \quad (17)$$

Нетрудно видеть, что функционал (17) квадратично зависит от коэффициентов a_j, поэтому всю теорию МНК можно переписать для этого случая.

$$\bar{\varphi} = \begin{bmatrix} \varphi_0(x_1) & \varphi_1(x_1) & \cdots & \varphi_k(x_1) \\ \varphi_0(x_2) & \varphi_1(x_2) & \cdots & \varphi_k(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_0(x_n) & \varphi_1(x_n) & \cdots & \varphi_k(x_n) \\ \bar{\varphi}_{(0)} & \bar{\varphi}_{(1)} & & \bar{\varphi}_{(k)} \end{bmatrix}$$

Если столбцы расширенной матрицы данных $\bar{\varphi}$ линейно независимы, то решение задачи (16) является решением линейной системы уравнений:

$$\bar{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \begin{bmatrix} (\bar{\varphi}_{(0)}, \bar{\varphi}_{(0)}) & \cdots & (\bar{\varphi}_{(0)}, \bar{\varphi}_{(k)}) \\ \vdots & \ddots & \vdots \\ (\bar{\varphi}_{(k)}, \bar{\varphi}_{(0)}) & \cdots & (\bar{\varphi}_{(k)}, \bar{\varphi}_{(k)}) \end{bmatrix} \times \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} (\bar{\varphi}_{(0)}, \bar{y}) \\ \vdots \\ (\bar{\varphi}_{(k)}, \bar{y}) \end{bmatrix} \qquad (19)$$

Если обобщенный полином (14) построен, то для анализа качества приближения используются те же методы, что для МНК - полиномов.

Пример: x_i, y_i, i=1,n – известно, что зависимость y(x) – линейная и проходит через начало координат. В этом случае расширенная матрица данных имеет 1 столбец.

$\Rightarrow$ Нормальная система уравнений сводится к одному линейному уравнению.

Пример: Изучая экономику США 20-века нетрудно заметить, что выпуск промышленной продукции I (в $) зависит от стоимости основных фордом K (в $) и стоимости рабочей силы L (в $). Эта зависимость имеет вид: $y = \alpha K^{\beta} L^{\gamma}$, $\alpha, \beta, \gamma - \textit{параметры}$ $(20*)$

$$\ln y = \ln\alpha + \beta \ln K + \gamma \ln L \quad (20)$$

Преобразуем исходные данные K_i, L_i, y_i и переходим к данным: $\hat{y}_i = \ln y_i$, $K_i = \ln K_i$, $L_i = \ln L_i$ нетрудно видеть, что (2) является обобщенным полиномом:

$$\hat{y} = \alpha + \beta K + \gamma L$$

$$a_0 = \alpha,\ \varphi_0 = 1; \qquad a_1 = \beta, \varphi_1 = K; \qquad a_2 = \gamma, \varphi_2 = L$$

формула (20*) – это производственная формула Кобба-Дугласа.

Стараются преобразовать данные так, чтобы зависимость для преобразованных данных стала линейной или полиномиальной.

Пример: Пенициллиновый бульон:

Результаты эксперимента содержат значительное влияние случайных факторов, таким образом, зависимость достаточно точно подбирать не нужно. $\hat{y} = a_0 + a_1 x$, $\hat{y} = a_0 + a_1 x + a_2 x^2$

8.2 Решение задачи Коши для ДУ методами Рунге-Кутта.

Правило Рунге: предположим, что некоторую величину F можно сосчитать несколькими способами. $F_{пр}^{(1)}$ *и* $F_{пр}^{(2)}$.

Пусть $F = \underbrace{F_{пр}^{(1)}}_{\textit{погр. выч. 1-м способом}} + \underbrace{Ah^p}_{\textit{гл.член погрешности}} + o(h^p)$ (1) $A = const \neq 0$, $h - \textit{малая величина}$

$$F = \underbrace{F_{пр}^{(2)}}_{\textit{погр. выч. 2-м способом}} + A\left(\frac{h}{2}\right)^p + o(h^p),\ p > 0 - \textit{малая величина}\ (2)$$

$(1)=(2)$ $F_{пр}^{(1)}+Ah^p+o(h^p)=F_{пр}^{(2)}+A\left(\frac{h}{2}\right)^p+o(h^p)$ $o(h^p)$ – в разных частях равенства различны, но по правилу Рунге ими можно пренебречь. $\boxed{F_{пр}^{(1)}+Ah^p=F_{пр}^{(2)}+A\left(\frac{h}{2}\right)^p}$ (3), на самом деле здесь приближенное равенство, но это не важно, (3) позволяет оценить главный член погрешности на основе результатов 2-х приближенных вычислений.

$F_{пр}^{(2)}-F_{пр}^{(1)}=Ah^p\left(1-\left(\frac{1}{2}\right)^p\right)\Rightarrow\boxed{Ah^p=\frac{F_{пр}^{(2)}-F_{пр}^{(1)}}{1-\left(\frac{1}{2}\right)^p}}$ (4) эта формула дает оценку главного члена погрешности для первого способа вычисления по правилу Рунге. Аналогично получается оценка для главного члена погрешности 2-го способа вычисления $A\left(\frac{h}{2}\right)^p$ (5). Оценив главный член погрешности можно уточнить значение F:

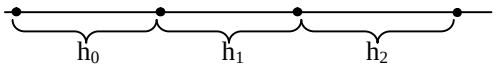
$$\boxed{F_{уточ}\overset{def}{=}F_{пр}^{(1)}+Ah^p=F_{пр}^{(1)}+\frac{F_{пр}^{(2)}-F_{пр}^{(1)}}{1-\left(\frac{1}{2}\right)^p}}\quad(6)$$

<u>Замечание:</u> Методы Рунге-Кутта можно пользоваться тогда, когда коэффициент A достаточно большое по сравнению с коэффициентом главного члена разложения $o(h^p)$.
Пример: $100h+0.1h^2$ – ok.
$10^{-6}h+10^{12}h^2$ – здесь правило Рунге не работает.

Лекция №36. Применение методов Рунге-Кутта.

Общие сведения о решении задачи Коши.

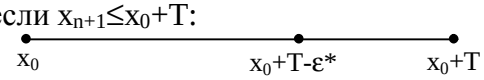
$\begin{cases}\frac{du}{dx}=f(x,u)\\ u(x_0)=u_0\end{cases}$ (1) – Найти решение задачи (1): траекторию проходящую на плоскости (x,u) через точки (x_0,u_0).
(x_i, u_i), i=0,1,2,… - точки лежащие на точном решении задачи (1)
(x_i, v_i), i=0,1,2,… - точки соответствующие приближенному решению задачи (1). $(x_0,v_0)=(x_0,u_0)$
Приближенную траекторию можно считать с постоянным или переменным шагом.
$h_n=x_{n+1}-x_n$



$h\overset{def}{=}x_{n+1}-x_n$ - шаг интегрирования при переходе от x_n к x_{n+1}.

Задачу можно рассматривать на ограниченном отрезке $x\in[x_0,x_0+T]$ или применить подход Адамса – считать до тех пор, пока не будет обнаружена какая-то закономерность (устойчивое решение, ∞, ассимптотичность).

Если задача ставиться на отрезке, то используют параметр ε* - критерий точности при выходе на правую границу.
ε*: пусть (x_n, v_n) вычисляют точку (x_{n+1}, v_{n+1}) в том случае, если $x_{n+1}\le x_0+T$:
Если: $x_0+T-\varepsilon^*\le x_{n+1}\le x_0+T$,



то после вычисления (x_{n+1}, v_{n+1}), подсчет заканчивается.

Метод Эйлера.

(2) $v_0=u_0$;
$x_{n+1}=x_n+h$ (h либо постоянен, либо меняется);
$v_{n+1}=v_n+hf(x_n,v_n)$ 1768 год.

Пример: $(1)\begin{cases}\dfrac{du}{dx}=5u \\ u(x_0)=u_0\end{cases}$ $\quad v_0=u_0$, $\quad x_{n+1}=x_n+h$, $\quad v_{n+1}=v_n+5hv_n=v_n(1+5h),\ n=0,1,2,\ldots$

Обоснование метода Эйлера:

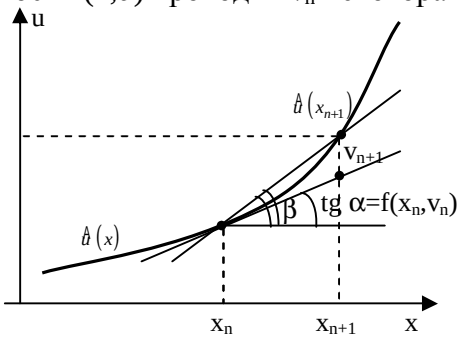
$\dfrac{u_{n+1}-u_n}{x_{n+1}-x_n}\approx\dfrac{du}{dx}(x_n)-$ аппроксимирует 1-ю производную функции $u(x)$ в точке x_n $\qquad \dfrac{u_{n+1}-u_n}{x_{n+1}-x_n}=f(x_n,u_n)\ (1^*)$

Поскольку точное решение задачи (1*) не совпадает с точным решением (1), то вместо u используют v: $v_{n+1}=v_n+(x_{n+1}-x_n)f(x_n,v_n)$.

геометрический смысл:

Рассмотрим задачу (1), пусть u(x) – траектория, которая проходила через точку (x_0, u_0) ⇒ f(x*, u*) это u'(x*) – тангенс угла наклона L касательной к этой траектории в точке (x*,u*).

Пусть (x_n,v_n) - точка, посчитанная численным методом. На плоскости (x,u) проходит v_n некоторая траектория задачи:



$$\begin{cases}\dfrac{du}{dx}=f(x,u) \\ u(x_n)=v_n\end{cases}\quad (3)$$

$\hat{u}(x)-$ *траектория*

$\hat{u}(x_{n+1})-$ *точка на решении* $\hat{u}(x)$ *и соответственно* x_{n+1}

tg β - f(x_n,v_n) – где β угол наклона секущей,

проходящей через точки: (x_n,v_n) *и* $\left(x_{n+1},\hat{u}(x_{n+1})\right)$(обе точки на точной траектории системы)

Очевидно, что: $\dfrac{\hat{u}(x_{n+1})-v_n}{x_{n+1}-x_n}\overset{def}{=}tg\beta\ (4)$. В методе Эйлера $\dfrac{v_{n+1}-v_n}{x_{n+1}-x_n}=f(x_n,v_n)\ (4^*)$.

Сравнив (4) и (4*) видно, что погрешность метода Эйлера в том, что он заменяет угол наклона секущей на угол наклона касательной.

В методе Эйлера, точки расположенные на точном решении связаны соотношением (4), а точки, расположенные на численном решении – соответствуют (4*).

Методы Рунге-Кутта отличаются от метода Эйлера тем, что аппроксимируют tg β более точным способом.

Расчетные формулы методов Рунге-Кутта.

$$\begin{cases}\dfrac{du}{dx}=f(x,u) \\ u(x_0)=u_0\end{cases}\quad (1)$$

Метод 2-го порядка:

1. $\begin{cases}x_{n+1}=x_n+h \quad v_0=u_0 \\ v_{n+1}=v_n+h\cdot f\left(x_n+\dfrac{h}{2},x_n+\dfrac{h}{2}\cdot f(x_n,v_n)\right)\end{cases}$

2. $\begin{cases}x_{n+1}=x_n+h, \qquad v_0=u_0 \\ v_{n+1}=v_n+\dfrac{h}{2}\left[f(x_n,v_n)+f\left(x_n+h,v_n+hf(x_n,v_n)\right)\right]\end{cases}$

N Q Q Q Q Q Q Q O Q Q Q Q Q Q Q P

$\approx tg\beta$

Методы 4-го порядка (Мак-Кракен, Дорн).

$$1.\ \begin{cases} x_{n+1} = x_n + h \quad v_0 = u_0 \\ v_{n+1} = v_n + \dfrac{h}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right) \\ k_1 = f\left(x_n, v_n\right);\ k_2 = f\left(x_n + \dfrac{h}{2}, v_n + \dfrac{h}{2}k_1\right);\ k_3 = f\left(x_n + \dfrac{h}{2}, v_n + \dfrac{h}{2}k_2\right);\ k_4 = f\left(x_n + h, v_n + hk_3\right) \end{cases}$$

$$2.\ v_{n+1} = v_n + h\left(p_1 k_1(h) + p_2 k_2(h) + \ldots + p_q k_q(h)\right)$$

q – количество стадий; $p_1, p_2, \ldots, p_q$ – заданные числа $k_1(h)=f(x_n,v_n)$

$$k_q(h) = f\left(x_n + \alpha_q h, v_n + h\sum_{i=1}^{q-1} \beta_{qi} k_i(h)\right)$$

Параметры $\alpha_1,\alpha_2,\ldots,\alpha_q$ и β_{ji} – в справочнике.

Табличная запись метода Рунге-Кутта.

№ стадии	p_*	α_*	β_{**}
1	p_1	-	-
2	p_2	α_2	β_{21}
3	p_3	α_3	β_{31} β_{32}
…	…	…	…
q	p_q	α_q	β_{q1} β_{q2} … β_{qq-1}

Запишем метод Рунге-Кутта 4-го порядка в табличном виде:

№ стадии	p_*	α_*	β_{**}
1	1/6	-	-
2	1/3	½	½
3	1/3	½	0,1/2
4	1/6	1	0,0,1

Пример: применение методов Рунге-Кутта.

$$\begin{cases} u' = 5u \\ u(x_0) = u_0 \end{cases} \quad \textit{применим метод I.}$$

$$v_0 = u_0,\ x_{n+1} = x_n + h; \qquad v_{n+1} = v_n + h \cdot f(\ldots,\ldots); \qquad f(x_n, v_n) = 5v_n$$

$$v_n + \frac{h}{2} f(x_n, v_n) = v_n\left(1 + 5\frac{h}{2}\right);\ v_{n+1} = v_n + h \cdot 5v_n\left(1 + \frac{5}{2}h\right) = v_n\left(1 + 5h\left(1 + \frac{5}{2}h\right)\right) = v_n\left(1 + 5h + \frac{5 \cdot 5}{2}h^2\right)$$

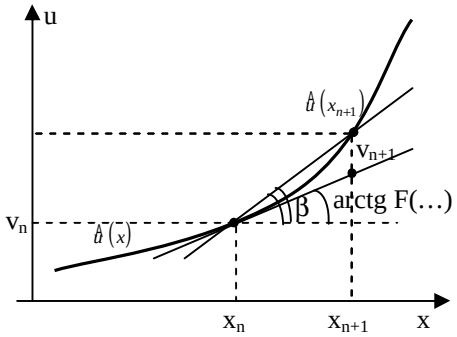
Порядок методов Рунге-Кутта.

Порядок методов Рунге-Кутта удобно определить на основе геометрических представлений.

$$\begin{cases} \dfrac{du}{dx} = f(x,u) \\ u(x_0) = u_0 \end{cases}$$ пусть вычислена (x_n,v_n), рассмотрим задачу (3):

$$\begin{cases} \dfrac{du}{dx} = f(x,u) \\ u(x_n) = v_n \end{cases} \quad \hat{u}(x) - \textit{точное решение } (3).$$

$\hat{u}(x_{n+1}) - \textit{точка на } \hat{u}(x)$



Есть соотношение (4), связывающее точки на $\hat{u}(x)$: $\dfrac{\hat{u}(x_{n+1})-v_n}{x_{n+1}-x_n}=tg\beta$

Метод РК общего вида: $v_{n+1}=v_n+h\cdot F(x_n,v_n,h)$ (5) $\Rightarrow \dfrac{v_{n+1}-v_n}{x_{n+1}-x_n}=F(x_n,v_n,h)$ – аппрокс. tg β.

Определение:

Метод РК (5) имеет порядок p, если $F(x_n, v_n, h)-tg\beta = 0(h^p)$ – величина порядка p.

Лекция №37.

Погрешность методов Рунге-Кутта. Связь порядка и погрешности с переменным шагом.

$$(1)\begin{cases}\dfrac{du}{dx}=f(x,u)\\ u(x_0)=u_0\end{cases}\quad \begin{cases}v_0=u_0\ ,x_{n+1}=x_n+h\\ v_{n+1}=v_n+h\cdot F(x_n,v_n,h)\end{cases},\ где\ F(x_n,v_n,h)=\sum_{i=1}^{q}p_i\cdot k_i(h)$$

$(x_i,v_i),\ i=0,1,2,\ldots$ – *точки метода* $(u_i,u_i), i=0,1,2,\ldots$ – *точки точного решения* (1)

Определение:

Глобальной погрешностью метода Рунге-Кутта в точке x_{n+1} называется $E_{n+1}=u_{n+1}-v_{n+1}$, где (6) u_{n+1} – точное решение (1) в точке x_{n+1}.

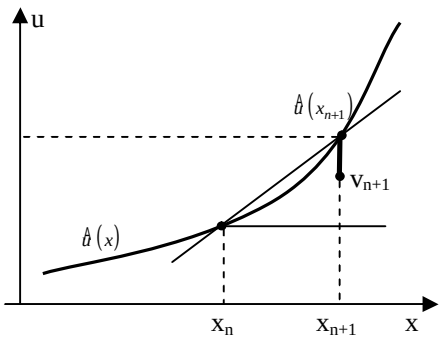
Определение:

Локальной погрешностью метода Рунге-Кутта называют отрыв численного метода от траектории, проходившей через предыдущую точку: $\begin{cases}\dfrac{du}{dx}=f(x,u)\\ u(x_n)=v_n\end{cases}$ $\hat{u}(x)$ – *точное решение* (3).

Локальная погрешность это: $e_{m+1}=\hat{u}(x_{n+1})-v_{n+1}$ (7)

В процессе счета, глобальную погрешность контролировать нельзя, а локальную можно контролировать.

Утверждение: Если метод Рунге-Кутта имеет порядок p+1: $e_{n+1}=0(h^{p+1})$.



Доказательство:

$$e_{n+1}=\hat{u}(x_{n+1})-v_{n+1}=\hat{u}(x_{n+1})-(v_n+hF(x_n,v_n,h))=$$

$$(\hat{u}(x_{n+1})-v_n)-hF(x_n,v_m,h)\boxed{=}$$

$$\hat{u}(x_{n+1})-v_n=(x_{n+1}-x_n)tg\beta$$

$$\boxed{=}\ h\underbrace{(tg\beta-F(x_n,v_n,h))}_{0(h^p)-т.к.\ метод\ порядка\ p(def)}=h\cdot 0(h^p)=0(h^{p+1})$$

Следствие: $e_{n+1}=\underbrace{A(x_n,v_n)\cdot h^{p+1}}_{гл.\ член\ лок.\ погрешности}+o(h^{p+1})$

Построение методов РК заданного порядка.

Чтобы метод РК имел порядок p, необходимо и достаточно, чтобы разложение $\sum_{i=1}^{q}p_ik_i(h)$ по степени h, отличалось от разложения tg β по степени h на величину $0(h^p)$.

$$u(x_{n+1})=u(x_n)+hu'(x_n)+\frac{h^2}{2}u''(x_n)+\ldots+\frac{h^s}{s!}u^{(s)}(x_n)+\frac{h^{s+1}}{(s+1)!}u^{(s+1)}(\xi),\ \xi\in[x_n,x_{n+1}]$$

$$=u(x_n)+h\cdot\left(u'(x_n)+\frac{h}{2}u''(x_n)+\ldots+\frac{h^{s-1}}{s!}u^{(s)}(x_n)+\frac{h^s}{(s+1)!}u^{(s+1)}(\xi)\right),\ \xi\in[x_n,x_{n+1}]$$

Если u(x) является решением уравнения: $(3)\dfrac{du}{dx} = f(x,u), u(x_n) = v_n$

$$\Rightarrow u'(x_n) = f(x_n, v_n);\quad u''(x) = \frac{d}{dx} f(x,u) = f'_x + f'_u \cdot \frac{du}{dx} = f'_x + f'_u \cdot f$$

$$u''(x_n) = [f'_x + f'_u \cdot f]\Big|_{\substack{x=x_n \\ u=u_n}};\qquad u^{(l)}(x_n) - \textit{можно выразить через } f \textit{ в точке } x_n, v_n$$

$$u'''(x) = \frac{d}{dx}(f'_x + f'_u \cdot f) = (f''_{xx} + f''_{xu} \cdot f) + \ldots$$

Если u(x) – решение (3), то $u(x_{n+1}) = \hat{u}(x_{n+1}),\ u(x_n) = \hat{u}(x_n) = v_n$

$$\boxed{\frac{u(x_{n+1}) - v_n}{h} \overset{def}{=} tg\beta = \left(f(x_n, v_n) + \frac{h}{2}(f'_x + f'_u \cdot f)\Big|_{\substack{x_n \\ v_n}} + \ldots + 0(h^s) \right) \quad (9)}$$

- это разложение tg по степени h. Для того, чтобы строить методы РК заданного порядка, нужно уметь строить разложение функции 2-х переменных в ряд Тейлора.

$$f(x,u) = f(x^*,u^*) + (x - x^*) f'_x(x^*,u^*) + (u - u^*) f'_u(x^*,u^*) + o((x - x^*) + (u - u^*)) \quad (10)$$

Пример: рассмотрим РК: $I.\quad v_{n+1} = v_n + h\, \underbrace{f\left(x_n + \frac{h}{2}, v_n + \frac{h}{2} f(x_n v_n)\right)}_{F(x_n, v_n, h)}$

$$F(x_n, v_n, h) = f(x_n, v_n) + \frac{h}{2} f'_x(x_n, v_n) + \frac{h}{2} f(x_n, v_n) f'_u(x_n, v_n) + o\left(\frac{h}{2} + \frac{h}{2} f\right) \quad (11)$$

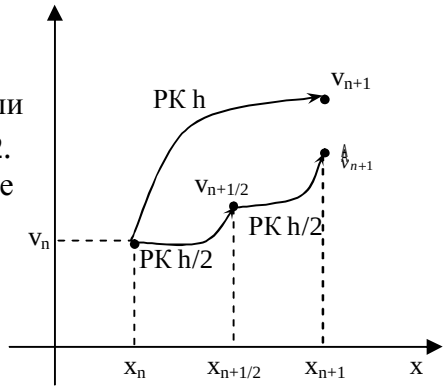
Сравним (9) и (11):

Если в формуле (9), в качестве s взять s=2 ⇒ (9) и (11) отличаются на слагаемое 2-го порядка ⇒ метод РК I имеет 2-й порядок.

Д/з: проверить РК II.

Метод Рунге-Кутта с контролем погрешности на шаге.

Метод имеет порядок p для контрля локальной погрешности используем число ε, которое задаем сами. Если разница между v_{n+1} и $\hat{v}_{n+1}$ маленькая, то v_{n+1} принимают в качестве решения. Если разница большая, то повторяем все то же самое из x_n с шагом h/2. Если разница слишком маленькая, то v_n – принимается в качестве решения и работаем с шагом в 2 раза большим.



(x_{n+1}, v_{n+1}) – *метод РК из* (x_n, v_n) *с h.*

$(x_{n+1/2}, v_{n+1/2})$ – *метод РК из* (x_n, v_n) *с* $h/2$.

$(x_{n+1}, \hat{v}_{n+1})$ – *метод РК из* $(x_{n+1/2}, v_{n+1/2})$ *с* $h/2$.

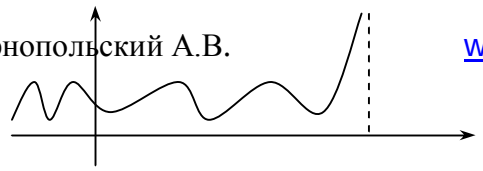
Затем вычисляем: $S = \dfrac{\hat{v}_{n+1} - v_{n+1}}{2^p - 1} (12)$

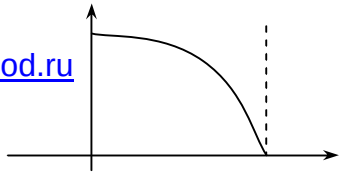
1) $\dfrac{\varepsilon}{2^{p+1}} \le |S| \le \varepsilon$, то (x_{n+1}, v_{n+1}) – решение.

2) $|S| > \varepsilon$, то делим h пополам и повторяем расчет из (x_n, v_n).

3) $|S| < \dfrac{\varepsilon}{2^{p+1}}$, то (x_{n+1}, v_{n+1}) – принято, но h увеличиваем вдвое.

Пример: случаи, в которых нужно менять шаг.





Связь метода (12) и локальной погрешности.

Утверждение о форме представления погрешности.

$e_{n+1} \overset{def}{=} \hat{u}(x_{n+1}) - v_{n+1} = A(x_n, v_n)h^{p+1} + o(h^{p+1})$ – локальная погрешность действий с шагом h.

$$e_{n+1/2} \overset{def}{=} \hat{u}(x_{n+1/2}) - v_{n+1/2} = A(x_n, v_n)\left(\frac{h}{2}\right)^{P+1} + o\left(\left(\frac{h}{2}\right)^{P+1}\right)$$

Оказывается: рассмотрим

$$\begin{cases} \dfrac{du}{dx} = f(x,u) \\ u(x_{n+1/2}) = v_{n+1/2} \end{cases} \quad - \text{его решение } \hat{u}(x) \quad (13)$$

$$\hat{u}(x_{n+1}) - \hat{v}_{n+1} = \underbrace{\hat{u}(x_{n+1}) - \hat{u}(x_{n+1})}_{=\text{с разн. 2-х траекторий в (.) } x_{n+1}} + \underbrace{\hat{u}(x_{n+1}) - \hat{v}_{n+1}}_{\text{это локальная погрешность метода РК с 1/2 из промож. (.)}} = A(x_{n+1/2}, v_{n+1/2})\left(\frac{h}{2}\right)^{p+1} + o\left(\left(\frac{h}{2}\right)^{p+1}\right)$$

$$e_{n+1/2} = \hat{u}(x_{n+1/2}) - v_{n+1/2}$$

Оказывается :1) $\hat{e}_{n+1} = A(x_n, v_n)\left(\frac{h}{2}\right)^{p+1} + o\left(\left(\frac{h}{2}\right)^{p+1}\right)$ 2) $c = A(x_n v_n)\left(\frac{h}{2}\right)^{p+1} + o\left(\left(\frac{h}{2}\right)^{p+1}\right)$

Таким образом:

$$\hat{u}(x_{n+1}) - \hat{v}_{n+1} = 2A(x_n, v_n)\left(\frac{h}{2}\right)^{p+1} + o\left(\left(\frac{h}{2}\right)^{p+1}\right) \ (15); \quad \hat{u}(x_{n+1}) - v_{n+1} = A(x_n, v_n)h^{p+1} + o(h^p) \quad (14)$$

Будем рассуждать так же, как и при выводе правила Рунге.

$$\hat{v}_{n+1} - v_{n+1} = Ah^{p+1}\left(1 - \left(\frac{1}{2}\right)^p\right) \qquad Ah^{p+1} = \frac{\hat{v}_{n+1} - v_{n+1}}{2^p - 1}2^p \quad (16)$$

Утверждение:

При использовании метода Рунге-Кутта (5) с правилами контроля погрешности (12) погрешность будет изменяться по правилу: $e_{n+1} = \dfrac{\hat{v}_{n+1} - v_{n+1}}{2^p - 1}2^p + o(h^{p+1})$.

Следствие: $\dfrac{\varepsilon}{2^{p+1}} \le \dfrac{|e_{n+1}|}{2^p} \le \varepsilon 2^p \quad (17)$

Лекция №38.

$$(1)\begin{cases} \dfrac{du}{dx} = f(x,u) \\ u(x_0) = u_0 \end{cases} \quad (5)\begin{cases} v_0 = u_0 \\ x_{n+1} = v_n + h \\ v_{n+1} = v_n + h \cdot F(x_n, v_n, h) \end{cases} ; \quad F(x_n, v_n, h) = \sum_{i=1}^{q} p_i k_i(h); \quad \frac{\varepsilon}{2^{p+1}} \le \left|\frac{\hat{v}_{n+1} - v_{n+1}}{2^p - 1}\right| \le \varepsilon \quad (12)$$

Здесь v_{n+1} и $\hat{v}_{n+1}$ два приближенных решения для x_{n+1}.

Смысл правила (12): $\dfrac{\varepsilon}{2} \le |e_{n+1}| \le \varepsilon 2^p \ (17)$ - это предположение (не всегда правда)

$$e_{n+1} = A(x_n, v_n)h^{p+1} + o\left(h^{p+1}\right)\ (I); \quad A(x_n, v_n)h^{p+1}\ (II) - \textit{оценивается}\ \frac{\hat{v}_{n+1} - v_{n+1}}{2^p - 1}2^p\ (III)$$

Таким образом, (17) – похоже на правду, если (III) похоже на (II) и если (h^{p+1}) в (I) действительно мало.

Пример: метод РК II:

$$v_0 = u_0$$

$$x_{n+1} = x_n + h$$

$$v_{n+1} = v_n + \frac{h}{2}\left[f(x_n, v_n) + f\left(x_n + h, v_n + h \cdot f(x_n, v_n)\right)\right] \quad p = 2$$

Решающее правило: $\frac{\varepsilon}{2^3} \le \left|\frac{\hat{v}_{n+1} - v_{n+1}}{2^2 - 1}\right| \le \varepsilon; \quad e_{n+1} \approx \frac{\hat{v}_{n+1} - v_{n+1}}{3} \cdot 4 -$ позволяет оценивать локальную погрешность на шаге

Пример: РК порядка:

$$\frac{\varepsilon}{2^5} \le \left|\frac{\hat{v}_{n+1} - v_{n+1}}{2^4 - 1}\right| \le \varepsilon; \quad e_{n+1} \approx \frac{\hat{v}_{n+1} - v_{n+1}}{15} \cdot 16$$

Обсудим формулу (12):

нижняя граница: $\frac{\varepsilon}{2^{p+1}}$ верхняя граница: ε $S = \left|\frac{\hat{v}_{n+1} - v_{n+1}}{2^p - 1}\right|$. Допустим при

x_{n+1} $S = \varepsilon + \delta\ (\delta > 0) \Rightarrow x_{n+1}$ - не принимается, и процедура повторяется из (x_n,v_n) с шагом h/2.

при первой попытке $e_{n+1}^{(I)} = A(x_n, v_n)h^{p+1} + o\left(h^{p+1}\right)$

при второй попытке $e_{n+1}^{(II)} = A(x_n, v_n)\left(\frac{h}{2}\right)^{p+1} + o\left(h^{p+1}\right)$

$\frac{e_{n+1}^{(I)}}{e_{n+1/2}^{(II)}} = 2^{p+1} \to e_{n+1/2}^{(II)} - e_{n+1}^{(I)}/2^{p+1}$; Если $S^{(I)}=\varepsilon+\delta$, то $S^{(II)}=(\varepsilon+\delta)1/2^{p+1}-\varepsilon/2^{p+1}+\delta/2^{p+1}$

$\Rightarrow$ на второй попытке: $\frac{\varepsilon}{2^{p+1}} < \left|S^{(II)}\right| < \varepsilon$, таким образом 1 способ выбора контрольных границ связан с поведением локальной погрешности метода.

p- порядок метода: $F(x_n, v_n, h)-tg\beta=0(h^p)$

q – число стадий, количество слагаемых типа p_i, $k_i(h)$.

Если нужен метод РК p, то q>p!

Зачем нужны несколько методов Рунге-Кутта одного порядка: для некоторых f(x,u) подходят одни формулы, а для других – другие.

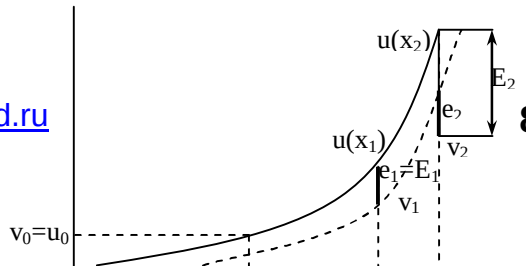
В методе с контролем шага(18) используется увеличение или деление шага на 2.

$$\alpha h, \text{ где } \alpha \text{ такое, чтобы: } \left|\frac{\hat{v}_{n+1} - v_{n+1}}{2^p - 1}\right| \approx \varepsilon,\ \textit{но} \le \varepsilon$$

Если метод РК порядка p $\Rightarrow$ его локальная погрешность имеет порядок p+1. Смысл числа p: при некоторых специальных предположениях по поводу уравнения (I) глобальная погрешность метода порядка p, тоже имеет порядок p.

$E_{n+1} = u_{n+1} - v_{n+1}; \quad u_{n+1} = u(x_{n+1}) - \textit{точное решение}\ (1); \quad v_{n+1} - \textit{значение метода РК при}\ x = x_{n+1}$

Теорема:



Пусть дана (1), где f(x,u) непрерывна по x,u и

$$\left|\frac{\partial f}{\partial u}\right| \le L \ при \ \forall x \in [x, x+T], \ \forall u \in (-\infty, +\infty).$$

Пусть для решения (1) на $[x_0, x_0+T]$ используется метод РК с постоянным шагом h=T/n, n – целое. Тогда $\max\limits_{i=0,n} |E_i| \le Mh^p$

Замечание: теорема справедлива не для всех уравнений вида (1). Справедлива, например, для линейных: $\frac{du}{dx} = \alpha x$

8.2.4 Счет с переменным шагом, с использованием контроля слагаемых.

Точка (x_n,v_n) вычислена; нужно найти (x_{n+1}, v_{n+1}) и оценить локальную погрешность:

$$e_{n+1} \overset{def}{=} \hat{u}(x_{n+1}) - v_{n+1}$$

Для этого использовался метод РК, применяемый с разным шагом.

Выясним, что получится, если применить метод РК порядка p РК порядка s, p<s.

$$PK(p): \tilde{v}_{n+1} = v_n + h \cdot F(x_n, v_n, h) \qquad PK(s): v_{n+1} = v_n + h \cdot F(x_n, v_n, h)$$

$$(18) \ PK(p): \tilde{e}_{n+1} = \hat{u}(x_{n+1}) - \tilde{v}_{n+1} = A(x_n, v_n) h^{p+1} + o(h^{p+1});$$

$$(19) \ PK(s): e_{n+1} = \hat{u}(x_{n+1}) - v_{n+1} = A(x_n, v_n) h^{s+1} + o(h^{s+1}); \quad s < p$$

Из (18) и (19) выразим $\hat{u}(x_{n+1})$: $\hat{u}(x_{n+1}) = Ah^{p+1} + o(h^{p+1}) + \tilde{v}_{n+1} = Ah^{s+1} + o(h^{s+1}) + v_{n+1}$

$$(20) \ \tilde{v}_{n+1} - v_{n+1} = \underbrace{Ah^{s+1}}_{глав.\ слагаемое} + \underbrace{o(h^{s+1}) - A(h^{p+1}) - o(h^{p+1})}_{=o(h^{s+1})}; \quad (21) \ Ah^{s+1} = \tilde{v}_{n+1} - v_{n+1} + o(h^{s+1})$$

При решение задачи (1) методами РК(s) и (p) где s<p, локальную погрешность методом порядка s можно оценить величиной: $e_{n+1} \approx \hat{v}_{n+1} - v_{n+1} = 0(h^{s+1})$

Пример: рассмотрим РК 4-го порядка и РК(2).

$$e_{n+1} \approx \tilde{v}_{n+1} - v_{n+1} = s^* - контрольное\ слагаемое$$

$$\tilde{v}_{n+1} = v_n + \frac{h}{6}(k_1 + 2k_2 + 2k_2 + k_4) \ (23); \quad v_{n+1} = v_n + hf\left(x_n + \frac{h}{2}, v_n + \frac{h}{2} f(x_n, v_n)\right) (24)$$

Очевидно, что $\tilde{v}_{n+1} = v_{n+1} + hk_2$ формулы (23) и (24) совпадают после подстановки коэффициентов.

$$e_{n+1} = h\left(\frac{k_1}{6} + \left(\frac{2}{6} - 1\right)k_2 + \frac{2}{6}k_3 + \frac{1}{6}k_4\right) = 0(h^3)$$

правило контроля погрешности на шаге можно построить аналогично.

Пусть метод РК порядка p и РК порядка s таковы, что коэффициенты k_i на первых стадиях обоих методов совпадают:

$$v_{n+1} = v_n + h\sum_{i=1}^{q} p_i k_i(h); \ k_i(h) = \tilde{k}_i(h), i = 1, r; \qquad \tilde{v}_{n+1} = v_n + h\sum_{i=1}^{q} p_i \tilde{k}_i(h) \quad r = \min(q_s, \tilde{q}_p)$$

⇒ счет с переменным шагом с использованием контрольного слагаемого позволяет значительно сэкономить количество вычислений функции F.

Лекция №39. #8.2.5 Итоги.

Пример: метод Рунге-Кутта-Мерсона.

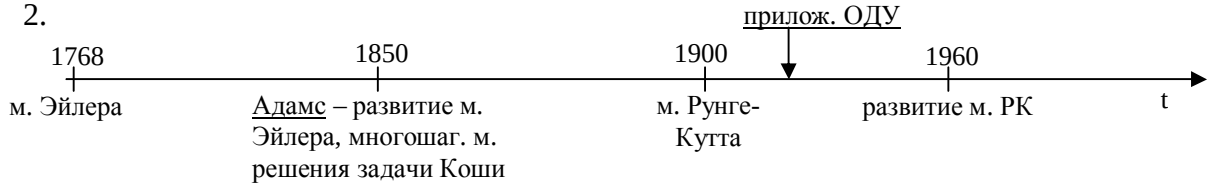
Для счета используем формулу 3-го порядка, а для контроля погрешности 4-го порядка $s^* \approx e_{n+1}$ (это локальная погрешность метода 3-го порядка) – имеет порядок (3+1)=4.

$\frac{\varepsilon}{2^4} = \frac{\varepsilon}{2^{p+1}} \le |S^*| \le \varepsilon$ – неравенство для построения решающего правила.

Какое использовать ε неважно, важно, во сколько раз отличается верхняя и нижняя граница.

Замечания по поводу решения задачи Коши методами Рунге-Кутта:

1. Правило Рунге нужно выучить как самостоятельное утверждение. Обычно в ЧМ оно напрямую негде не используется, оно используется как некоторые рассуждения. Мы пользовались этим способом, когда строили метод с контролем погрешности на шаге.
2.



Различие методов Рунге-Кутта и многошаговых методов.

Методы решения задачи Коши делят на 2 класса:

- многошаговые методы;
- методы Рунге-Кутта.

Пусть x_n, v_n – вычислена ⇒ в методе РК при вычислении x_{n+1}, v_{n+1} приходиться несколько раз вычислять правую часть уравнения: f(x,u), при этом, для подсчета точки x_{n+1}, v_{n+1} точки, предшествующие x_n, v_n не используются.

В многошаговых методах для вычисления x_n, v_n используются $(x_{n-1}, v_{n-1}),\ldots,(x_{n-k}, v_{n-k})$. При этом говорят, что многошаговый метод имеет шаговость k+1. От предшествующих точек требуется, чтобы x, были на одинаковых расстоянии: $x_{n+1} - x_n = x_n - x_{n-1} = x_{n-1} - x_{n-2} = \ldots = x_{n-k+1} - x_{n-k}$.

Основным недостатком методов РК является необходимость много раз вычислять функцию, а достоинством многошаговых – возможность сэкономить на вычислении функций, потому, что мы используем предыдущие точки.

С другой стороны методы РК легко программируются, и у них легко менять шаг. Для многошаговых методов нужны специальные методы для досчета точек, если в ходе решения ДУ нужно изменить шаг.

В течении длительного времени методы РК не использовались, т.к. считались трудоемкими из-за вычисления функций. С появлением вычислительной техники РК стали самыми популярными.

3. Решение системы ОДУ задачи Коши:

Методы Рунге-Кутта нетрудно переписать применительно к решению системы уравнений, но бывают жесткие системы ДУ, для которых нежелательно применять метод РК.

Систему ДУ называют жесткой, если решение системы содержит быстро затухающие и медленно затухающие компоненты. К тому моменту, когда быстро затухающая компонента решения задачи Коши на самом деле затухнет будет применяться метод с большим шагом интегрирования (в методе РК это практически невозможно).

Были построены специальные неявные многошаговые методы для решения жестких систем (пр. метод Гира)

8.3 Метод Бубнова – Галеркина.

Метод решения краевых задач, линейных и нелинейных, но обязательно с линейными граничными условиями.

Методы решения краевых задач:

1) Разностная схема, построенная на основе разностных аналогов физических законов сохранения.
2) Методы вариационного типа: решение сводиться к решению задачи оптимизации некоторого функционала в бесконечно мерном пространстве, которая затем решается приближенно в некотором конечномерном пространстве: (метод Ритца).
3) Проекционные методы (метод коллокаций, метод наименьших квадратов, метод Галеркина). В любом из этих методов решение краевой задачи, т.е. отыскание элемента некоторого бесконечномерного пространства, сводится к отысканию функции в конечномерном пространстве.
   - метод аппроксимаций функционалов: решение краевой задачи сводиться к решению задачи оптимизации в бесконечномерном пространстве для некоторого функционала, затем этот функционал аппроксимируют некоторым функционалом из конечномерного пространства и решают конечномерную задачу оптимизации.

Бесконечномерная задача сводиться к конечномерной.
Метод Бубнова-Галеркина работает даже тогда, когда для него ничего не доказано.

8.3.1 Переход от ДУ с неоднородными граничными условиями к ДУ с однородными граничными условиями.

$$(1)\begin{cases} u''+xu'+u=4x-1 \\ u(0)=0 \quad u(1)=0 \\ u(x), x\in[0,1] \end{cases} \qquad (2)\begin{cases} \omega''+x\omega'+\omega=2x \\ \omega(0)=1,\ \omega(1)=0 \\ \omega(x)\in[0,1] \end{cases}$$

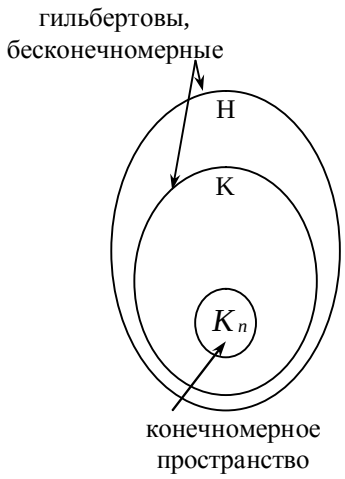
однородные гр. условия неоднородные граничные условия

пусть η(x): η(0)=1, η(1)=0 решение задачи (2) будем искать в виде ω(x)=u(x)+η(x) (3)
уравнение для u(x): u(x)=ω(x)-η(x) ⇒ (u”+η”)+x(u’+η’)+(u+η)=2x

$$(4)\begin{cases} u''+xu'+u=2x-(\eta''+x\eta'+\eta) \\ u(0)=0,\ u(1)=0 \end{cases}$$

получена краевая задача с однородными граничными условиями, например η(x)=1-x.

8.3.2 Постановка задачи при использовании метода Бубнова-Галеркина.



1) Краевая задача должна быть сведена к задаче с однородными граничными условиями – это позволяет искать решение в виде разложения по базису.
Пусть L – линейный дифференциальный оператор с областью определения K и областью значений H.
l – оператор, обеспечивающий однородность граничных условий.

$$(5)\begin{cases} Lu=f \\ lu=0 \end{cases}$$ - линейная краевая задача в абстрактной форме.

K – бесконечномерное пространство, включенное в H, которое в некоторой другой норме можно рассматривать как самостоятельное гильбертово пространство, и которое нужно, чтобы корректно определить область определения некоторого линейного дифференциального оператора, присутствующего в краевой задаче.
Таким образом, L –дифференциальный оператор с областью определения K и областью значений H. W_2^s – соболевское пространство (s – отвечает за гладкость). K_n - конечномерное пространство, элементы которого являются элементами K, размерность n и для каждой функции $\forall\varphi\in K_n$ lφ=0.
<u>Идея метода:</u>

Пусть $\varphi_1,...,\varphi_n$ – базис пространства K_n, приближенное решение задачи (5) будем искать в виде разложения по этому базису: $v = \alpha_1\varphi_1 + \alpha_2\varphi_2 + ... + \alpha_n\varphi_n$ (6)

Коэффициенты α_I ищут из условия ортогональности невязки уравнения (5) на решении v всем базисным функциям φ_i, i=1,n.