Московский государственный университет
имени М. В. Ломоносова

Механико-математический факультет

Конспект курса
Численные методы

Осенний семестр 2022, 4-ый курс, эконом. поток.

Выполнил:
студент 4 курса 432 группы
Шерстобитов Андрей Сергеевич

Лектор:
Доктор физико-математических наук,
профессор кафедры теории вероятностей
Корнев Андрей Алексеевич

Москва
2023

Содержание

1 Вычислительная погрешность. Устойчивость задачи и численного алгоритма.

Определим вычислительную погрешность и машинную точность. Для этого нам понадобятся основы машинной арифметики.

Наиболее распространенная форма представления действительных чисел в компьютерах - числа с плавающей точкой. Множество F чисел с плавающей точкой характеризуется четырьмя параметрами: основанием системы счисления p, разрядностью t и интервалом показателей $[L;U]$. Каждое число x, принадлежащее F, представимо в виде

$$x = \pm\left(\frac{d_1}{p} + \frac{d_2}{p^2} + \ldots + \frac{d_t}{p^t}\right)p^{\alpha};\ 0 \le d_i \le p-1;\ i = 1,\ldots,t; L \le \alpha \le U$$

d_i называют *разрядами*, t - *длиной мантиссы*, α - *порядком числа*. *Мантиссой* (дробной частью) x называют число в скобках. Множество F называют *нормализованным* если $\forall\ x \neq 0 \Rightarrow d_1 \neq 0$.

Пример 1.1. $p = 10,\ t = 6$

$$x = 1723,4835 = +\left(\frac{1}{10} + \frac{7}{10^2} + \frac{2}{10^3} + \frac{3}{10^4} + \frac{4}{10^5} + \frac{8}{10^6}\right)10^4 = 0,172348 \cdot 10^4$$

Округление чисел при работе на компьютере с точностью ε это некоторое отображение $float : \mathbb{R} \to F$ удовлетворяющее условию:

$$\forall\ y \in \mathbb{R},\ float(y) \in F,\ float(y) \neq 0 \to float(y) = y \cdot (1+\eta), |\eta| < \varepsilon$$

Таким образом относительная погрешность

$$\left|\frac{float(y) - y}{float(y)}\right| = \left|\frac{y \cdot (1+\eta) - y}{y}\right| = |\eta| \le \varepsilon \equiv \text{const}$$

Величина ε имеет порядок p^{1-t}. Величину ε часто называют *машинной точностью*.

В современных компьютерах число с *плавающей точкой* 'double' занимает 8 байт: 1 бит под знак, 11 битов под экспоненту, оставшиеся 52 под мантиссу. Диапазон значений: $2^{\pm 2^{10}} \simeq 10^{\pm 308}$, машинная точность: $2^{-53} \simeq 10^{-16}$.

Алгоритмы, традиционно применяемые в точной арифметике, могут некорректно работать при расчетах из-за конечной точности на ЭВМ. Как следствие, к методам и постановкам задач вычислительной математики предъявляют дополнительные требования.

1. Решаемая численно задача должна быть устойчива, т.е. малое изменение входных параметров не должно значительно менять результат;

   Пример 1.2 (Неустойчивая задача)**.** Рассотрим возмущенную матрицу Уилкинсона

   $$A(\varepsilon) = \begin{pmatrix} 20 & 20 & & & 0 \\ & 19 & 20 & & \\ & & 18 & \ddots & \\ & & & \ddots & \\ 0 & & & & 20 \\ \varepsilon & 0 & & & 1 \end{pmatrix}$$

   Определитель этой матрицы (считаем по столбцам) равен

   $$\det(A(\varepsilon)) = 20! - \varepsilon \cdot 20^{19}$$

   Характеристический многочлен матрицы

   $$\det(A(\varepsilon) - \lambda I) = (20-\lambda)\cdots(1-\lambda) - 20^{19}\varepsilon$$

   При $\varepsilon = 0$ $\det(A(0)) = 20!$, а минимальное собственное число $\lambda_{min} = 1$. Рассмотрим $\varepsilon = 20^{-19} \cdot 20! \approx 5 \cdot 10^{-7}$:

   $$\det(A(5 \cdot 10^{-7})) \approx 0$$

   То есть при достаточно малом ε определитель матрицы изменился на 20!, а величина наименьшего собственного значения стала равна 0.

   Таким образом вычисление определителя является неустойчивой задачей, так как *незначительная погрешность во входных данных может существенно исказить ответ.*

2. Выбранный алгоритм должен быть численно устойчив, т.е. ошибки округления в промежуточных вычислениях не должны искажать окончательный ответ;

   Пример 1.3 (Численно неустойчивый алгоритм)**.** Вычисляется сумма $\sum_{i=1}^{10^3} 1/i^2$. Какой алгоритм даст большую точность?

   $$S_0 = 0;\ S_n = S_{n-1} + \frac{1}{n^2},\ n = 1, \ldots, 10^3$$

   или

   $$R_{10^3+1} = 0;\ R_{n-1} = R_n + \frac{1}{n^2},\ n = 10^3, \ldots, 1$$

   Рассмотрим более простой пример: в компьютерной арифметике надо сложить два числа при $t = 5, p = 10$:

   $$float(100,01) + float(0,0001) = 0,10001 \cdot 10^3 + 0,10000 \cdot 10^{-3} = float(100,0101) = 0,10001 \cdot 10^3$$

   То есть сложение чисел разного порядка может привести к потере точности.

   Возвращаясь к задаче: следует воспользоваться вторым способом. При вычислении первым способом происходит потеря точности в результате сложения чисел S_{n-1} и $1/n^2$, существенно отличающихся по величине.

3. Имеющиеся вычислительные ресурсы (память, быстродействие, программное обеспечение) должны позволить реализовать алгоритм и получить ответ за требуемое время.

   Пример 1.4 (Непозволительно долгий алгоритм)**.** Метод Крамера решения систем линейных алгебраических уравнений с невырожденной матрицей $n \times n$ позволяет найти точное решение, вычислив $(n+1)$ определитель матриц размерности $n \times n$. Оценим $T(n)$ - время работы алгоритма.

   Вычисление одного определителя методом миноров реализуется за $\sim nn!$ арифметических действий, для вычисления $(n+1)$ определителя потребуется $N \sim n(n+1)!$ арифметических действий. Например, для $n = 20, 100$ имеем:

   $$20! \approx 2.4 \cdot 10^{18},\ N(20) \approx 10^{21}; 100! \approx 10^{158},\ N(100) \approx 10^{162}$$

   На современных компьютерах для выполнения $N(20)$ операций потребуется $3 \cdot 10^{139}$ лет.

2 Линейные разностные уравнения n-го порядка. Теоремы о представлении общего решения однородного уравнения и общего решения неоднородного уравнения.

Опр. 2.1. Пусть неизвестная функция y, заданные функции a, f – функции одного целочисленного аргумента k.

$$a_0(k)y(k) + a_1(k)y(k+1) + \ldots + a_{n-1}(k)y(k+n-1) + a_n(k)y(k+n) = f(k) \Leftrightarrow Ly = f$$

Тогда уравнение при $f(k),\ a_0(k),\ a_n(k) \neq 0$ называется неоднородным линейным разностным уравнением n-го порядка. Порядок – количество начальных условий, при которых уравнение является разрешимым. Уравнение $Ly = 0$ называется однородным.

Пример 2.1.

$$S(n) = \sum_{i=0}^{n}(1 + i + i^3) \underset{\text{переход к разностному уравнению}}{\Rightarrow} S_{n+1} = S_n + 1 + (n+1) + (n+1)^3$$

Получили уравнение первого порядка, в качестве начальных условий возьмем $S_0 = +1+(0+1)+(0+1)^3 = 3$.

Теорема 2.1. *Пусть $y^{(1)}(k), \ldots, y^{(n)}(k)$ - произвольные линейно независимые решения линейного однородного разностного уравнения n-го порядка $Ly = 0$, тогда общее решение можно представить в виде*

$$y(k) = \sum_{i=1}^{n} c_i y^{(i)}(k)$$

c_i в данной записи порождаются из начальных условий задачи.

Доказательство. 1. Покажем, что если функция совпадает на n точках, то она совпадает и на всех остальных.

Так как наше уравнение имеет порядок n, то $\exists\ a_0(k), \ldots, a_{n-1}(k) \neq 0$.

Перепишем исходное уравнение в двух видах

$$y(k+n) = -\sum_{i=0}^{n-1} \frac{a_i(k)}{a_n(k)} y(k+i) \tag{1}$$

$$y(k) = -\sum_{i=1}^{n} \frac{a_i(k)}{a_0(k)} y(k+i) \tag{2}$$

Таким образом, если мы знаем n точек $y(k_0), \ldots, y(k_0 + n - 1)$, то из равенства (1) мы можем восстановить $y(k)\ \forall\ k \geq k_0 + n$, а из равенства (2) мы можем восстановить $y(k)\ \forall\ k \leq k_0$. То есть если мы имеем n точек, то мы можем однозначно восстановить все решение, а это значит, что если два решения совпадают на n точках, то они тождественно равны $\forall\ k$.

2. Дано $\left\{y^{(i)}(k)\right\}_{i=1}^{n}$ – n линейно независимых решений нашего уравнения. Зафиксировав $k_0, \ldots, k_{n-1}$, мы получим базис в пространстве $\mathbb{R}^n$. В этом базисе мы можем выразить искомое решение как линейную комбинацию элементов базиса

$$y(k) = \sum_{i=1}^{n} c_i y^{(i)}(k),\ k = k_0, \ldots, k_{n-1}$$

Из предыдущего пункта знаем, что если решение совпадает на n точках, то совпадает и везде, то есть

$$y(k) = \sum_{i=1}^{n} c_i y^{(i)}(k),\ \forall k$$

□

Опр. 2.2. *Общим* решением задачи называют то решение, которое можно получить из любых начальных условий.

Теорема 2.2. *Пусть $y^o(k)$ - общее решение однородной задачи $Ly = 0$. Пусть $y^1(k)$ - некоторое частное решение неоднородной задачи $Ly = f$. Тогда любое решение неоднородной задачи можно представить в виде*

$$y(k) = y^o(k) + y^1(k)$$

Доказательство. Пусть $y(k)$ – какое-либо решение задачи $Ly = f$, $y^1(k)$ – некоторое частное решение этой же задачи. Тогда $y(k) - y^1(k)$ является решением задачи $Ly = 0$:

$$L(y(k) - y^1(k)) = L(y(k)) - L(y^1(k)) = f(k) - f(k) = 0 \Rightarrow y(k) = y^o(k) + y^1(k)$$

□

3 Линейные разностные уравнения n-го порядка с постоянными коэффициентами. Формулировка теорем о представлении общего решения однородного уравнения и частного решения неоднородного уравнения с квазимногочленом в правой части. Форма записи действительного решения.

Опр. 3.1. Рассматриваем $a_i(k) \stackrel{\text{def}}{=} a_i \equiv \text{const}\ \forall i = 0, \dots, n.\ a_0 \neq 0,\ a_n \neq 0$. Для удобства обозначим $y(k) \stackrel{\text{def}}{=} y_k$ и $f(k) \stackrel{\text{def}}{=} f_k$. Тогда линейным разностным уравнением с постоянными коэффициентами n-го порядка называют

$$a_0 y_k + a_1 y_{k+1} + \ldots + a_{n-1} y_{k+n-1} + a_n y_{k+n} = f_k \Leftrightarrow Ly = f$$

Для однозначного определения решения требуется задать n условий, например, $y_i = b_i,\ i = 0, \ldots n-1$.

Аналогично обычным дифференциальным уравнениям с постоянными коэффициентами будем искать решение однородного разностного уравнения в виде $y_k = \mu^k$.

После подстановки этого выражения в разностное уравнение и сокращения на μ^k получим *характеристический многочлен.*

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

Утверждение. *Пусть $\mu_1, \ldots, \mu_r$ - различные корни характеристического многочлена, а $\sigma_1, \ldots, \sigma_r$ - их кратности ($\sum \sigma_i = n$). Тогда общее решение однородного уравнения с постоянными коэффициентами n-го порядка можно представить в виде*

$$\begin{array}{rccccccc} y_k = & c_{11}\mu_1^k & + & c_{12}k\mu_1^k & + & \ldots & + & c_{1\sigma_1}k^{\sigma_1-1}\mu_1^k \\ + & c_{21}\mu_2^k & + & c_{22}k\mu_2^k & + & \ldots & + & c_{2\sigma_2}k^{\sigma_2-1}\mu_2^k \\ + & \ldots & & \ldots & & \ldots & & \ldots \\ + & c_{r1}\mu_r^k & + & c_{r2}k\mu_r^k & + & \ldots & + & c_{r\sigma_r}k^{\sigma_r-1}\mu_r^k \end{array}$$

где c_{ij} - произвольные постоянные.

Доказательство. Без доказательства. □

Пример 3.1. Найти общее решение уравнения

$$by_{k+1} - cy_k + ay_{k-1} = 0$$

Найдем корни характеристического многочлена

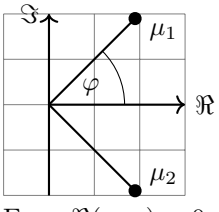
$$P(\mu) = b\mu^2 - c\mu + a = 0 \Rightarrow \mu_{1,2} = \frac{c \pm \sqrt{D}}{2b},\ D = c^2 - 4ab$$

Решение зависит от значения дискриминанта D:

1. $D > 0$: $\mu_1 \neq \mu_2 \in \mathbb{R} \Rightarrow y(k) = C_1\mu_1^k + C_2\mu_2^k$
2. $D = 0$: $\mu_1 = \mu_2 = \mu \in \mathbb{R} \Rightarrow y(k) = C_1\mu^k + C_2 k\mu^k$
3. $D < 0$: Так как $a, b, c \in \mathbb{R}$, то $\mu_{1,2} = \rho \exp(\pm i\varphi)$ - комлексно-сопряженные.

$$\Re(\mu_{1,2}) = \frac{c}{2b},\ \Im(\mu_{1,2}) = \pm\frac{\sqrt{|D|}}{2b}$$

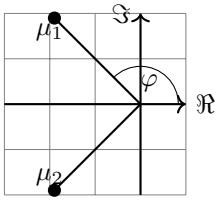
(a) Если $\Re(\mu_{1,2}) > 0$:



$$\rho = \sqrt{\left(-\frac{c}{2b}\right)^2 + \left(\frac{\sqrt{|D|}}{2b}\right)^2} = \sqrt{\frac{c^2}{4b^2} + \frac{4ab - c^2}{4b^2}} = \sqrt{\frac{a}{b}}$$

$$\tan\varphi = \frac{\Im(\mu_{1,2})}{\Re(\mu_{1,2})} \Rightarrow \varphi = \arctan\frac{\sqrt{|D|}}{c}$$

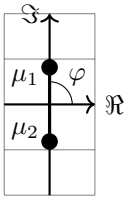
(b) Если $\Re(\mu_{1,2}) < 0$:



$$\rho = \sqrt{\frac{a}{b}}$$

$$\tan\varphi = \frac{\Im(\mu_{1,2})}{\Re(\mu_{1,2})} \Rightarrow \varphi = \pi - \arctan\frac{\sqrt{|D|}}{c}$$

(c) Если $\Re(\mu_{1,2}) = 0 \Rightarrow c = 0$:



$$\rho = \sqrt{\frac{a}{b}}$$
$$\varphi = \frac{\pi}{2}$$

Представив $\mu_{1,2} \in \mathbb{C} = \rho(\cos\varphi \pm i\sin\varphi)$ и подставив в $y_k = c_1\mu_1^k + c_2\mu_2^k$ получим *форму записи действительного решения*

$$y_k = \rho^k(\tilde{C}_1\cos k\varphi + \tilde{C}_2\sin k\varphi)$$

Как и в случае дифференциальных уравнений, частное решение разностного уравнения для правой части специального вида может быть найдено методом неопределенных коэффициентов

Утверждение. *Если правая часть задачи с постоянными коэффициентами $Ly = f$ принимает вид квазимногочлена*

$$f_k = \alpha^k(P_{m_1}(k)\cos(\varphi k) + Q_{m_2}(k)\sin(\varphi k))$$

где m_1 и m_2 степени соответсвующих полиномов, то частное решение может принимать вид

$$y_k^1 = \alpha^k k^s(\tilde{P}_{\tilde{m}}\cos(\varphi k) + \tilde{Q}_{\tilde{m}}\sin(\varphi k))$$

где $\tilde{m} = \min(m_1, m_2)$, $s = 0$, если $\alpha\exp(\pm i\varphi)$ не является корнем характеристического многочлена, иначе s - его кратность.

Доказательство. Без доказательства. □

Чтобы найти коэффициенты многочленов $\tilde{P}_{\tilde{m}}$ и $\tilde{Q}_{\tilde{m}}$, надо подставить частное решение в неоднородное уравнение и приравнять коэффициенты при подобных членах.

Пример 3.2. Найти вид частного решения уравнения

$$y_{k+2} + y_k = \cos\frac{\pi}{2}k$$

- Корни характеристического уравнения:$P(\mu) = \mu^2 + 1 = 0 \Rightarrow \mu_{1,2} = \pm i$
- $\cos\frac{\pi}{2}k$ является квазимногочленом с $\alpha = 1,\ m_1 = m_2 = 0,\ \varphi = \frac{\pi}{2}$.
- $\alpha\exp(\pm i\varphi) = \exp\left(\pm i\frac{\pi}{2}\right) = \pm i$ - корень характеристического многочлена кратности $1 \Rightarrow s = 1$

Вид частного решения принимает вид

$$y_k^1 = k\left(c_1\cos\frac{\pi}{2}k + c_2\sin\frac{\pi}{2}k\right)$$

4 Фундаментальное решение разностного уравнения. Теорема о представлении частного решения неоднородного уравнения первого порядка с постоянными коэффициентами.

Рассматриваем неоднородное разностное уравнение n-го порядка $Ly = f$ с постоянными коэффициентами. Хотим построить частное решение для произвольной правой части.

Опр. 4.1. Фундаментальным решением G_k называют решение следующего разностного уравнения

$$a_0 y_k + a_1 y_{k+1} + \ldots + a_n y_{k+n} = \delta_k^0 = \begin{cases} 1, k = 0 \\ 0, k \neq 0 \end{cases}$$

Среди бесконечного количества решений нас будут интересовать ограниченные, то есть $|G_k| \leq \text{const } \forall\ k$.

Основная идея: так как не умеем строить для произвольного f_k, хотим научиться строить фундаментальное решение. Тогда мы сможем разбить f_k на бесконечное количество фундаментальных решений и воспользоваться тем, что $Ly_1 = f_1,\ Ly_2 = f_2 \Rightarrow L(y_1 + y_2) = f_1 + f_2$.

Пример 4.1. Найдем ограниченное фундаментальное решение задачи

$$ay_k + by_{k+1} = \delta^0,\ a, b \neq 0$$

Так как задача неоднородная, то решение можно представить в виде $y_k = y_k^o + y_k^1$. Найдем общее решение задачи. Для этого решим $ay_k + by_{k+1} = 0$

$$P(\mu) = a + b\mu = 0 \Rightarrow \mu = -\frac{a}{b} \Rightarrow y_k = C\left(-\frac{a}{b}\right)^k$$

Частное решение будем строить с помощью хитрого трюка. Запишем нашу задачу в виде системы

$$\begin{cases} ay_k + by_{k+1} = 0,\ k \leq -1 \\ ay_k + by_{k+1} = 1,\ k = 0 \Leftrightarrow ay_0 + by_1 = 1 \\ ay_k + by_{k+1} = 0,\ k \geq 1 \end{cases}$$

Так как знаем решение однородной задачи, то сразу же можем выписать решения для первого и третьего уравнения $y_k = C^-\left(-\frac{a}{b}\right)^k,\ k \leq -1$, $y_k = C^+\left(-\frac{a}{b}\right)^k,\ k \geq 1$. Обратим внимание, что константы C^- и C^+ разные, так как это две разных части одного решения! Осталось их найти.

Так как мы ищем любое частное решение задачи, то ради удобства возьмем $C^- = 0 \Rightarrow y_k \equiv 0\ \forall\ k \leq -1$.

Как найти y_0? Подставим в первое уравнение $k = -1$, тогда $ay_{-1} + by_0 = 0$. Так как при $k \leq -1$ $y_k \equiv 0$, то $by_0 = 0 \Rightarrow y_0 = 0$.

Теперь мы можем найти y_1: $ay_0 + by_1 = 1 \Rightarrow y_1 = \frac{1}{b}$.

Для того чтобы найти C^+ подставим в известное нам общее решение y_1:

$$y_1 = C^+\left(-\frac{a}{b}\right)^1 \Leftrightarrow \frac{1}{b} = C^+\left(-\frac{a}{b}\right) \Rightarrow C^+ = -\frac{1}{a}$$

Таким образом искомое решение принимает вид

$$y_k = C\left(-\frac{a}{b}\right)^k + \begin{cases} 0,\ k \leq 0 \\ -\frac{1}{a}\left(-\frac{a}{b}\right)^k,\ k \geq 1 \end{cases} = \begin{cases} C\left(-\frac{a}{b}\right)^k,\ k \leq 0 \\ \left(C - \frac{1}{a}\right)\left(-\frac{a}{b}\right)^k,\ k \geq 1 \end{cases}$$

Выделим из этого множества только ограниченные решения.

- Если $\left|\frac{a}{b}\right| > 1$, то первое уравнение системы будет ограничено при $k \to -\infty$. Второе уравнение наоборот будет стремиться к бесконечности поэтому C нужно взять $\frac{1}{a}$. $G_k = \begin{cases} \frac{1}{a}\left(-\frac{a}{b}\right)^k,\ k \leq 0 \\ 0,\ k \geq 1 \end{cases}$
- Если $\left|\frac{a}{b}\right| = 1$, то $\forall\ C$ решение будет ограниченным. $G_k = \begin{cases} C,\ k \leq 0 \\ C - \frac{1}{a},\ k \geq 1 \end{cases}$
- Если $\left|\frac{a}{b}\right| < 1$, то второе уравнение системы будет ограничено при $k \to +\infty$, тогда как второе будет неограниченно. Возьмем $C = 0$. $G_k = \begin{cases} 0,\ k \leq 0 \\ -\frac{1}{a}\left(-\frac{a}{b}\right)^k,\ k \geq 1 \end{cases}$

Теорема 4.1. *Пусть* $|G_k^n| \le \text{const}$, $|f_k| \le F = \text{const}$, $\left|\frac{a}{b}\right| \ne 1$. *Тогда ряд*

$$y_k = \sum_{-\infty}^{+\infty} f_n G_k^n$$

Будет абсолютно сходиться и являться решением неоднородной задачи $ay_k + by_{k+1} = f_k$.

Доказательство. • Пусть $\left|\frac{a}{b}\right| > 1$, тогда $G_k^n = \begin{cases} \frac{1}{a}\left(-\frac{a}{b}\right)^{k-n}, \ k-n \le 0 \\ 0, \ k-n \ge 1 \end{cases}$

$$\sum_{-\infty}^{+\infty} f_n G_k^n = \sum_{k-n\le 0} f_n \frac{1}{a}\left(-\frac{a}{b}\right)^{k-n} = \frac{1}{a}\sum_{n-k\ge 0} f_n \left(-\frac{b}{a}\right)^{n-k} \le \frac{|F|}{|a|}\sum_{n-k\ge 0}\underbrace{\left|\frac{b}{a}\right|^{n-k}}_{<1} = \frac{|F|}{|a|}\frac{1}{1-\left|\frac{b}{a}\right|} = \frac{|F|}{|b|-|a|}$$

Таким образом ряд сходится абсолютно, а значит возможна перестановка слагаемых. Проверим, что y_k действительно решение. Подставим в исходную задачу.

$$ay_k + by_{k+1} = a\left(\sum_{-\infty}^{+\infty} f_n G_k^n\right) + b\left(\sum_{-\infty}^{+\infty} f_n G_k^{n+1}\right) = \sum_{-\infty}^{+\infty} f_n(\underbrace{aG_k^n + bG_k^{n+1}}_{\delta_k^n}) = f_k$$

• Пусть $\left|\frac{a}{b}\right| < 1$, тогда $G_k^n = \begin{cases} 0, \ k \le 0 \\ -\frac{1}{a}\left(-\frac{a}{b}\right)^{k-n}, \ k-n \ge 1 \end{cases}$

$$\sum_{-\infty}^{+\infty} f_n G_k^n = \sum_{k-n\ge 1} f_n \frac{-1}{a}\left(-\frac{a}{b}\right)^{k-n} \le \frac{|F|}{|a|}\sum_{k-n\ge 1}\underbrace{\left|\frac{b}{a}\right|^{k-n}}_{<1} \le \frac{|F|}{|a|}\frac{1}{1-\left|\frac{a}{b}\right|} = |F|\frac{|b|}{|a|(|b|-|a|)}$$

Аналогично ряд сходится абсолютно.

□

Отметим, что изложенная техника применима для построения фундаментального решения для уравнения n-го порядка

5 Решение задач на собственные значения для разностных уравнений, сравнение с дифференциальным случаем.

Постановка задачи:

$$\begin{cases} \frac{y_{k+1}-2y_k+y_{k-1}}{h^2} = -\lambda y_k,\ h = \frac{2}{2N-1},\ 1 \le k \le N-1 \\ y_0 = 0;\ y_N = y_{N-1} \end{cases}$$

1. Запишем канонический вид. Найдем коэффициенты для краевых условий

$$k = 1 : \frac{y_2 - 2y_1}{h^2} = -\lambda y_1$$
$$k = N-1 : \frac{y_N - 2y_{N-1} + y_{N-2}}{h^2} = \frac{-y_{N-1} + y_{N-2}}{h^2} = -\lambda y_{N-1}$$

   Таким образом задачу можно переписать в матричном виде:

$$\begin{pmatrix} \frac{-2}{h^2} & \frac{1}{h^2} & & & & 0 \\ \frac{1}{h^2} & \frac{-2}{h^2} & \frac{1}{h^2} & & & \\ & \cdots & \cdots & & & \\ & & & \frac{1}{h^2} & \frac{-2}{h^2} & \frac{1}{h^2} \\ 0 & & & & \frac{1}{h^2} & \frac{-1}{h^2} \end{pmatrix} \begin{pmatrix} y_1 \\ \\ \vdots \\ \\ y_{N-1} \end{pmatrix} = -\lambda \begin{pmatrix} y_1 \\ \\ \vdots \\ \\ y_{N-1} \end{pmatrix}$$

2. Для более удобного решения сделаем замену $p = 1 - h^2\frac{\lambda}{2}$ и перепишем условие:

$$\begin{cases} y_{k+1} - 2py_k + y_{k-1} = 0,\ 1 \le k \le N-1 \\ y_0 = 0;\ y_N = y_{N-1} \end{cases}$$

   Решим полученную разностную задачу:

$$P(\mu) = \mu^2 - 2p\mu + 1 = 0 \Leftrightarrow \mu_{1,2} = p \pm \sqrt{p^2 - 1}$$

   Также по теореме Виета:

$$\mu_1 \cdot \mu_2 = 1 \tag{1}$$
$$\frac{\mu_1 + \mu_2}{2} = p \tag{2}$$

   (a) $\mu_1 \neq \mu_2$. Тогда общее решение имеет вид

$$y_k = C_1\mu_1^k + C_2\mu_2^k$$

   Подставим в начальные условия, чтобы найти C_1 и C_2:

$$\begin{cases} y_0 = 0 = C_1 + C_2 \Rightarrow C_2 = -C_1 \\ y_N = y_{N-1} : C_1\mu_1^N + C_2\mu_2^N = C_1\mu_1^{N-1} + C_2\mu_2^{N-1} \end{cases}$$

   Преобразуем второе равенство, используя первое:

$$C_1(\mu_1^N - \mu_2^N) = C_1(\mu_1^{N-1} - \mu_2^{N-1})$$

   Если $C_1 = 0$, то $C_2 = 0$, то $y_k \equiv 0$, что нам неинтересно, так как нулевой вектор не является собственным. Иначе

$$\mu_1^N - \mu_2^N = \mu_1^{N-1} - \mu_2^{N-1} \Leftrightarrow \mu_1^N - \mu_1^{N-1} = \mu_2^N - \mu_2^{N-1} \Leftrightarrow \mu_1^{N-1}(\mu_1 - 1) = \mu_2^{N-1}(\mu_2 - 1)$$

   Используем п.1 из теоремы Виета:

$$\mu_1^{N-1}(\mu_1 - \mu_1\mu_2) = \mu_2^{N-1}(\mu_2 - 1) \Leftrightarrow -\mu_1^N(\mu_2 - 1) = \mu_2^{N-1}(\mu_2 - 1)$$

   Заметим, что $\mu_2 \neq 1$, так как по теореме Виета $\mu_1 = \mu_2 = 1$ – противоречие.

$$\frac{\mu_1^N}{\mu_2^{N-1}} = -1 \Leftrightarrow \mu_1^{2N-1} = -1$$

Возьмем $2N-1$ комплексный корень из 1 и получим:

$$\begin{cases}\mu_1^{(m)} = \exp(\frac{\pi(2m+1)i}{2N-1}) \\ \mu_2^{(m)} = \exp(-\frac{\pi(2m+1)i}{2N-1})\end{cases} \quad m = 0, ..., 2N-2$$

Решение имеет вид:

$$y_k = C_1\mu_1^k + C_2\mu_2^k = C\left(\exp\left(\frac{\pi(2m+1)i}{2N-1}\right) - \exp\left(-\frac{\pi(2m+1)i}{2N-1}\right)\right) =$$
$$= 2C\left(\frac{\exp\left(\frac{\pi(2m+1)i}{2N-1}\right) - \exp\left(-\frac{\pi(2m+1)i}{2N-1}\right)}{2}\right) = C\sin\frac{\pi(2m+1)k}{2N-1}$$
$$m = 0, ..., 2N-2,\ k = 1, ..., N-1$$

Или иначе:

$$y_k = C\sin\frac{\pi(2m-1)k}{2N-1},\ m = 1, ..., 2N-1,\ k = 1, ..., N-1$$

Найдем собственные значения. По теореме Виета:

$$p = \frac{\mu_1+\mu_2}{2} = \frac{\exp\left(\frac{\pi(2m-1)i}{2N-1}\right) + \exp\left(-\frac{\pi(2m-1)i}{2N-1}\right)}{2} = \cos\left(\frac{\pi(2m-1)}{2N-1}\right)$$
$$m = 1, ..., 2N-1$$

$$p = 1 - \lambda\frac{h^2}{2} \Leftrightarrow$$
$$\lambda = \frac{2}{h^2}\left(1 - \cos\left(\frac{\pi(2m-1)}{2N-1}\right)\right) = \frac{4}{h^2}\sin^2\left(\frac{\pi(2m-1)}{2(2N-1)}\right)$$
$$m = 1, ..., 2N-1$$

У симметричной матрицы размера $N-1 \times N-1$ не может быть больше $N-1$ собственного значения, но выше мы получили $2N-1$. Посмотрим на них подробнее: обозначим $\alpha_m = \frac{\pi(2m-1)}{2(2N-1)}$. Тогда

$$\alpha_1 = \frac{2\pi}{2(2N-1)} - \frac{\pi}{2(2N-1)} = \frac{\pi}{2(2N-1)}$$
$$\alpha_2 = \frac{4\pi}{2(2N-1)} - \frac{\pi}{2(2N-1)} = 3\alpha_1$$
$$...$$
$$\alpha_{2N-1} = \frac{2\pi(2N-1)}{2(2N-1)} - \frac{\pi}{2(2N-1)} = \pi - \frac{\pi}{2(2N-1)} < \pi$$

То есть все $2N-1$ угол расположены в верхней части триногометрического круга. Но так как sin имеет одинаковые значения для I и II частей круга, то половина корней будет совпадать. Осталось посмотреть в какой части находится Nый корень:

$$\alpha_N = \frac{\pi(2N-1)}{2(2N-1)} = \pi/2 \Rightarrow \lambda = \frac{4}{h^2}$$

Но такое возможно тогда и только тогда, когда $\mu_1 = \mu_2$, что нам не подходит. Таким образом ответ:

$$y_k = C\sin\frac{\pi(2m-1)k}{2N-1} \qquad m = 1, ..., N-1$$
$$\lambda = \frac{4}{h^2}\sin^2\left(\frac{\pi(2m-1)}{2(2N-1)}\right) = (2N-1)^2\sin^2\left(\frac{\pi(2m-1)}{2(2N-1)}\right) \qquad k = 1, ..., N-1$$

(b) $\mu_1 = \mu_2$: из теоремы Виета следует $\mu_1 = \mu_2 = p$ и $\mu_1\mu_2 = 1 \Rightarrow \lambda = 0$ и $\lambda = \frac{4}{h^2}$. Вставим это решение в ответ из случая $\mu_1 \neq \mu_2$

Итоговый ответ:

$$y_k = C \sin \frac{\pi(2m-1)k}{2N-1} \qquad m = 1, ..., N-1$$

$$\lambda = \frac{4}{h^2} \sin^2 \left(\frac{\pi(2m-1)}{2(2N-1)} \right) \qquad k = 0, ..., N$$

6 Построение многочленов Чебышёва первого и второго рода.

Рассматривается рекурентное соотношение

$$y_{n+1}(x) = 2xy_n(x) - y_{n-1}(x),\ x = \text{const} \in \mathbb{R}$$

Так как x зафиксировано, то можем решить это соотношение как однородное разностное уравнение.

$$P(\mu) = \mu^2 - 2x\mu + 1 = 0 \Rightarrow \mu_{1,2} = x \pm \sqrt{x^2-1}$$

$$\begin{array}{ll} |x| \neq 1: & y_n(x) = C_1(x)(x+\sqrt{x^2-1})^n + C_2(x)(x-\sqrt{x^2-1})^n \\ |x| = 1: & y_n(x) = C_1(x) + C_2(x)n \end{array} \tag{1}$$

При $|x| < 1$ сделаем замену $x = \cos\varphi$ и получим тригонометрическую форму записи:

$$y_n(x) = \hat{C}_1(x)(\cos(n\arccos(x))) + \hat{C}_2(x)(\sin(n\arccos(x)))$$

Опр. 6.1. Многочленами Чебышёва первого рода называется последовательность многочленов, удовлетворяющих рекурентному соотношению

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x),\ T_0(x) = 1,\ T_1(x) = x$$

Теорема 6.1.

$$T_n(x) = \frac{(x+\sqrt{x^2-1})^n + (x-\sqrt{x^2-1})^n}{2}\ \forall x \in \mathbb{R}$$

Доказательство. Найдем C_1 и C_2 для первого уравнения системы (1)

$$\begin{cases} C_1(x+\sqrt{x^2-1})^0 + C_2(x-\sqrt{x^2-1})^0 = 1 \\ C_1(x+\sqrt{x^2-1})^1 + C_2(x-\sqrt{x^2-1})^1 = x \end{cases} \Rightarrow \begin{cases} C_1 + C_2 = 1 \\ (1-2C_2)\sqrt{x^2-1} = 0 \end{cases} \Rightarrow C_1 = C_2 = \frac{1}{2},\ x \neq \pm 1$$

Так как T_n порождается полиномами - непрерывными функциями, а предложенная в теореме запись - рациональная функция - непрерывна, то значит что в точках $x \pm 1$ из-за непрерывности они будут совпадать. □

Теорема 6.2.

$$T_n(x) = \cos(n\arccos(x)),\ |x| \leq 1$$

Доказательство. Найдем C_1 и C_2 для тригонометрической формы

$$\begin{cases} C_1(\cos(0\cdot\arccos(x))) + C_2(\sin(0\cdot\arccos(x))) = 1 \\ C_1(\cos(\arccos(x))) + C_2(\sin(\arccos(x))) = x \end{cases} \Rightarrow \begin{cases} C_1 = 1 \\ C_2(\sin(\arccos(x))) = 0 \end{cases} \Rightarrow \begin{cases} C_1 = 1 \\ C_2 = 0 \end{cases} \quad |x| \leq 1$$

По непрерывности аналогично получаем значения на краях интервала $(-1, 1)$. □

Опр. 6.2. Многочленами Чебышёва второго рода называется последовательность многочленов, удовлетворяющих рекурентному соотношению

$$U_{n+1}(x) = 2xU_n(x) - U_{n-1}(x),\ U_0(x) = 1,\ U_1(x) = 2x$$

Теорема 6.3.

$$U_n(x) = \frac{(x+\sqrt{x^2-1})^{n+1} - (x-\sqrt{x^2-1})^{n+1}}{2\sqrt{x^2-1}}\ \forall x \in \mathbb{R}$$

Доказательство. Уже знаем, что $\forall\ x \in \mathbb{R}$ решение можно представить в виде

$$U_n(x) = C_1(x)(x+\sqrt{x^2-1})^n + C_2(x)(x-\sqrt{x^2-1})^n$$

Осталось подобрать коэффициенты C_1 и C_2:

$$\begin{cases} C_1(x+\sqrt{x^2-1})^0 + C_2(x-\sqrt{x^2-1})^0 = 1 \\ C_1(x+\sqrt{x^2-1})^1 + C_2(x-\sqrt{x^2-1})^1 = 2x \end{cases} \Rightarrow \begin{cases} C_1 + C_2 = 1 \\ (1-2C_2)\sqrt{x^2-1} = x \end{cases} \Rightarrow$$

$$\Rightarrow \begin{cases} C_1 + C_2 = 1 \\ C_2 = \frac{\sqrt{x^2-1}-x}{2\sqrt{x^2-1}} \end{cases} \Rightarrow \begin{cases} C_1 = \frac{\sqrt{x^2-1}+x}{2\sqrt{x^2-1}} \\ C_2 = -\frac{x-\sqrt{x^2-1}}{2\sqrt{x^2-1}} \end{cases} \quad x \neq \pm 1$$

Аналогично из непрерывности следует тождественность на $x = \pm 1$ □

Теорема 6.4.

$$U_n(x) = \frac{\sin((n+1)\arccos x)}{\sin\arccos x},\ |x| \le 1$$

Доказательство. Сделаем замену $x = \cos\varphi$ в предыдущей теореме и воспользуемся формулой Муавра

$$\frac{(\cos\varphi + i\sin\varphi)^{n+1} - (\cos\varphi - i\sin\varphi)^{n+1}}{2i\sin\varphi} =$$

$$\frac{\cos((n+1)\varphi) + i\sin((n+1)\varphi) - \cos((n+1)\varphi) + i\sin((n+1)\varphi)}{2i\sin\varphi} = \frac{2i\sin((n+1)\varphi)}{2i\sin\varphi} = \frac{\sin((n+1)\arccos x)}{\sin\arccos x}$$

Аналогично из непрерывности получаем тождественность для $x = \pm 1$ □

Замечание. Обратим внимание, что $T_n(x)$ и $U_n(x)$ порождаются линейно независимыми комбинациями $(1, x)$ и $(1, 2x)$. Это значит, что любое рекурентное соотношение $y_{n+1}(x) = 2xy_n(x) - y_{n-1}(x)$ имеет решение $y_n(x) = C_1(x)T_n(x) + C_2(x)U_n(x)$.

7 Свойства многочленов Чебышёва первого рода: симметричность, нули, экстремумы.

Опр. 7.1. Многочленами Чебышёва первого рода называется последовательность многочленов, удовлетворяющих рекурентному соотношению

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x),\ T_0(x) = 1,\ T_1(x) = x$$

Теорема 7.1. *Старший коэффициент многочленов Чебышёва имеет вид* $T_n(x) = 2^{n-1}x^n + \dots$

Доказательство. База индукции $T_2 = 2x^2 - 1$ - верна. Шаг индукции:

$$T_{n-1}(x) = 2^{n-2}x^{n-1} + \dots \Rightarrow T_n(x) = 2xT_n(x) + T_{n-1}(x) = 2x \cdot (2^{n-2}x^{n-1} + \dots) + \dots = 2^{n-1}x^n + \dots$$

□

Теорема 7.2.

$$T_{2n}(-x) = T_{2n}(x),\ T_{2n+1}(-x) = -T_{2n+1}(x)$$

Доказательство. По индукции. □

Теорема 7.3. *Два различных многочлена Чебышёва ортогональны относительно скалярного произведения* $(T_n, T_m)_{p(x)}$*, где вес* $p(x) = \frac{1}{\sqrt{1-x^2}} > 0$ *п.в.*

$$\int_{-1}^{1} \frac{T_n(x)T_m(x)}{\sqrt{1-x^2}}dx = \begin{cases} 0,\ m \neq n \neq 0 \\ \frac{\pi}{2},\ m = n \neq 0 \\ \pi,\ m = n = 0 \end{cases}$$

Доказательство. Переходим к тригонометрической форме

$$\int_{-1}^{1} \frac{\cos(n\arccos(x))\cos(m\arccos(x))}{\sqrt{1-x^2}}dx = \left|\begin{smallmatrix} x=\cos\varphi \\ dx=-\sin\varphi d\varphi \\ 1\to 0 \\ -1\to\pi \end{smallmatrix}\right| = \int_0^{\pi} \frac{\cos(n\varphi)\cos(m\varphi)}{\sin\varphi}\sin\varphi d\varphi =$$

$$= \frac{1}{2}\int_0^{\pi} \cos((n+m)\varphi) + \cos((n-m)\varphi)d\varphi = \frac{\sin((n+m)\varphi)}{2(n+m)}\Big|_0^{\pi} + \frac{\sin((n-m)\varphi)}{2(n-m)}\Big|_0^{\pi} = \begin{cases} 0,\ m \neq n \neq 0 \\ \frac{\pi}{2},\ m = n \neq 0 \\ \pi,\ m = n = 0 \end{cases}$$

□

Теорема 7.4. $T_n(x)$ *на отрезке* $[-1, 1]$ *имеет* n *различных корней* $x_m = \cos\frac{(2m-1)\pi}{2n}$.

Доказательство. Найдем корни уравнения тригонометрической формы.

$$\cos(n\arccos(x_m)) = 0 \Leftrightarrow n\arccos(x_m) = -\frac{\pi}{2} + \pi m,\ m = 0, \pm1, \pm2, \dots$$

$$x_m = \cos\left(\frac{\pi(2m-1)}{2n}\right),\ m = 0, \pm1, \pm2, \dots$$

Получили много корней, а в теореме всего n. Посмотрим внимательнее на корни.

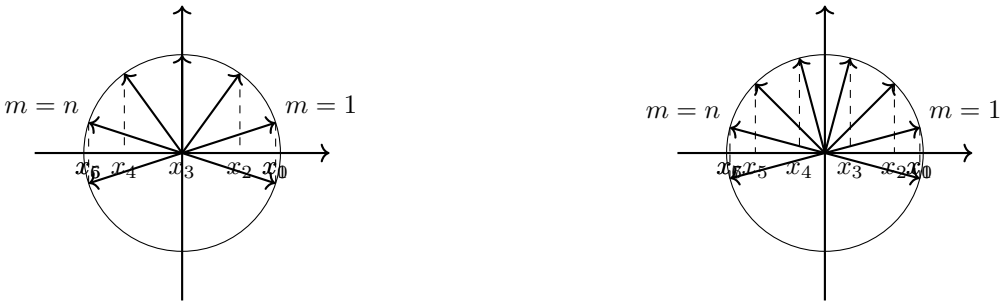


Рис. 1: Корни при $n = 5$ и $n = 6$

Из полученных множеств нужно выбрать только различные, а их как раз n. Обратим так же внимание, что при представлении в тригонометрической форме мы заложились, что $|x| \leq 1$, но так как на $[-1, 1]$ мы нашли все n корней, а у полинома n-ой степени больше быть не может, то мы нашли корни $\forall\ x \in \mathbb{R}$ □

Теорема 7.5. *На отрезке* $[-1,1]$ *имеется* $n+1$ *экстремум,* $T_n(x_{(m)}) = (-1)^m$. *Экстремумы имеют вид*

$$x_{(m)} = \cos\frac{\pi m}{n},\ m = 0,\dots,n$$

Доказательство. Найдем экрстремумы уравнения тригонометрической формы.

$$\cos(n\arccos(x_{(m)}) = \pm 1 \Leftrightarrow n\arccos(x_{(m)}) = \pi m,\ m = 0,\pm 1,\pm 2,\dots$$

$$x_{(m)} = \cos\frac{\pi m}{n},\ m = 0,\pm 1,\pm 2,\dots$$

Ищем различные экстремумы, аналогично корням, получаем $n+1$ штуку. □

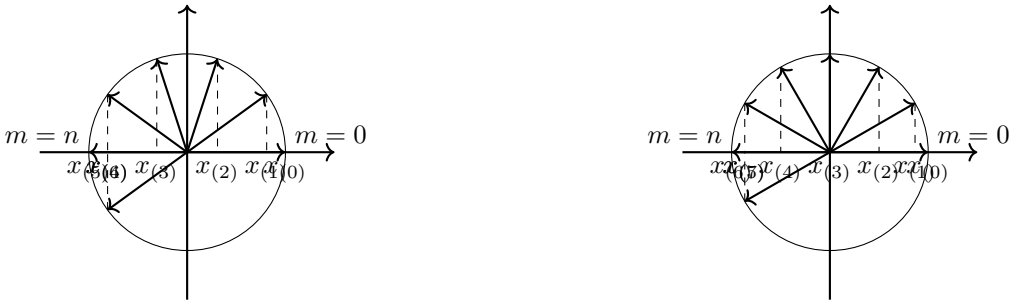


Рис. 2: Экстремумы при $n = 5$ и $n = 6$

Таким образом, мы можем сделать вывод о том как выглядят многочлены Чебышева.

Рис. 3: Многочлены Чебышёва степени $n = 5$ и $n = 6$

8 Экстремальные свойства многочленов Чебышёва первого рода на отрезке [a, b].

Теорема 8.1. *Среди всех многочленов со старшим коэффициентом* 1 *приведенный многочлен Чебышёва наименее отклоняется от нуля на отрезке* $[a;b]$. *Эквивалентная запись:*

$$\arg\left\{\inf_{P_n(x)=x^n+\dots}\max_{x\in[a,b]}|P_n(x)|\right\}=2^{1-n}\left(\frac{b-a}{2}\right)^n T_n\left(\frac{2x-(a+b)}{b-a}\right)=:\overline{T}_n(x)$$

Доказательство. 1. Перенесем на отрезок $[a,b]$ многочлен Чебышёва, используя замену

$$x=\frac{a+b}{2}+\frac{b-a}{2}t,\ t\in[-1,1]\Leftrightarrow t=\frac{2x-(a+b)}{b-a},\ x\in[a,b]$$

$$T_n\left(\frac{2x-(a+b)}{b-a}\right)\underset{\text{св-во}}{=}2^{n-1}\left(\frac{2x-(a+b)}{b-a}\right)^n+\dots$$

Нормируем для получения приведенного многочлена:

$$\overline{T}_n(x):=2^{1-n}\left(\frac{b-a}{2}\right)^n T_n\left(\frac{2x-(a+b)}{b-a}\right)=x^n+\dots$$

2. Пусть $\exists\ P_n^*(x),\ \|P_n^*(x)\|_{[a,b]}<\|\overline{T}_n(x)\|_{[a,b]}$. Рассмотрим $Q_{n-1}=\overline{T}_n(x)-P_n^*(x)$.

   - Многочлен $Q_{n-1}(x)\not\equiv 0$, так как $P_n^*(x)$ и $\overline{T}_n(x)$ имеют различные нормы.
   - Многочлен $Q_{n-1}(x)$ степени не больше $n-1$, так как старшие коэффициенты у $P_n^*(x)$ и $\overline{T}_n(x)$ равны и сократились.

3. Заметим, что $\operatorname{sgn}Q_{n-1}(x_{(m)})=\operatorname{sgn}\left(\overline{T}_n(x_{(m)})-P_n^*(x_{(m)})\right)=\operatorname{sgn}\overline{T}_n(x_{(m)})$, так как $\|P_n^*(x)\|_{[a,b]}<\|\overline{T}_n(x)\|_{[a,b]}$,

4. Знаем, что $\operatorname{sgn}\overline{T}_n(x_{(m)})=(-1)^m,\ m=0,\dots,n$. Значит изучаемый полином Q_{n-1} имеет $n+1$ экстремум, а значит имеет n корней, а значит $Q_{n-1}\equiv 0$. Противоречие.

□

Теорема 8.2. *Среди всех многочленов, которые в точке* $x_0=0$ *равны* 1, *приведенный многочлен Чебышёва наименее отклоняется от нуля на отрезке* $[a,b]$, *при условии, что* $x_0\notin[a,b]$.

$$\arg\left\{\inf_{P_n(x)=1+\dots}\max_{\substack{x\in[a,b]\\x_0\notin[a,b]}}|P_n(x)|\right\}=\frac{T_n\left(\frac{2x-(a+b)}{b-a}\right)}{T_n\left(\frac{-(a+b)}{b-a}\right)}=:\hat{T}_n$$

При этом если $0<a<b$:

$$\|\hat{T}_n\|=\frac{2}{q^n+q^{-n}}=\frac{2q^n}{q^{2n}+1}\le 2q^n,\ q=\frac{\sqrt{b}-\sqrt{a}}{\sqrt{b}+\sqrt{a}}<1$$

Доказательство. 1. Многочлен Чебышёва перенесенный на $[a,b]$ имеет вид

$$T_n\left(\frac{2x-(a+b)}{b-a}\right)=c_nx^n+\dots+c_1x_1+c_0$$

Тогда, чтобы найти значение c_0 возьмем многочлен Чебышёва в точке x_0:

$$T_n\left(\frac{-(a+b)}{b-a}\right)=c_n\cdot 0+\dots+c_1\cdot 0+c_0=c_0$$

Тогда приведенный многочлен с 1 в младшем коэффициенте имеет вид

$$\hat{T}_n=\frac{T_n\left(\frac{2x-(a+b)}{b-a}\right)}{T_n\left(\frac{-(a+b)}{b-a}\right)}$$

2. Пусть $\exists\ P_n^*(x),\ \|P_n^*(x)\|_{[a,b]}<\|\hat{T}_n(x)\|_{[a,b]}$. Рассмотрим $Q_n=\hat{T}_n(x)-P_n^*(x)$.

- Многочлен $Q_n(x) \not\equiv 0$, так как $P_n^*(x)$ и $\hat{T}_n(x)$ имеют различные нормы.
- Многочлен $Q_n(x)$ степени не больше n.

3. Аналогично предыдущей теореме $\operatorname{sgn} Q_n(x_{(m)}) = \operatorname{sgn} \hat{T}_n(x_{(m)}) = (-1)^m,\ m = 0, \ldots, n$. Значит $Q_n(x)$ имеет n корней и $n+1$ экстремум на $[a, b]$. Но заметим, что у $Q_n(x)$ нет свободного члена, так как они сократились. Это значит, что в точке $0 \notin [a, b]$ Q_n имеет еще один корень x_0. Значит $Q_n \equiv 0$. Противоречие.

   Заметим, что именно здесь важно, что $x_0 \notin [a, b]$, так как иначе нельзя сказать, что у Q_n имеется $n+1$ корень.

4. Посчитаем норму $\hat{T}_n(x)$ при $0 < a < b$:

$$\max_{x\in[a,b]} |\hat{T}_n(x)| = \max_{x\in[a,b]} \left| \frac{T_n\left(\overbrace{\frac{2x-(a+b)}{b-a}}^{t\in[-1,1]}\right)}{T_n\left(\frac{-(a+b)}{b-a}\right)} \right| = \frac{1}{\left|T_n\left(\frac{-(a+b)}{b-a}\right)\right|}$$

$$T_n\left(\frac{-(a+b)}{b-a}\right) = \frac{\left(\frac{-(a+b)}{b-a} + \sqrt{\left(\frac{-(a+b)}{b-a}\right)^2 - 1}\right)^n + \left(\frac{-(a+b)}{b-a} - \sqrt{\left(\frac{-(a+b)}{b-a}\right)^2 - 1}\right)^n}{2} = (\star)$$

$$\sqrt{\left(\frac{-(a+b)}{b-a}\right)^2 - 1} = \sqrt{\left(\frac{a+b}{a-b}\right)^2 - 1} = \sqrt{\frac{(a+b)^2 - (a-b)^2}{(a-b)^2}} = \sqrt{\frac{4ab}{(a-b)^2}} = \frac{2\sqrt{ab}}{b-a}$$

$$(\star) = \frac{\left(\frac{a+b}{a-b} + \frac{2\sqrt{ab}}{b-a}\right)^n + \left(\frac{a+b}{a-b} - \frac{2\sqrt{ab}}{b-a}\right)^n}{2} = \frac{\left(\frac{a+b}{a-b} - \frac{2\sqrt{ab}}{a-b}\right)^n + \left(\frac{a+b}{a-b} + \frac{2\sqrt{ab}}{a-b}\right)^n}{2} =$$

$$= \frac{\left(\frac{\sqrt{a}+\sqrt{b}}{\sqrt{a}-\sqrt{b}}\right)^n + \left(\frac{\sqrt{a}-\sqrt{b}}{\sqrt{a}+\sqrt{b}}\right)^n}{2} = \frac{(-q)^n + (-q)^{-n}}{2},\ q = \frac{\sqrt{b}-\sqrt{a}}{\sqrt{b}+\sqrt{a}}$$

□

9 Экстремальные свойства многочленов Чебышёва первого рода вне (a,b).

Теорема 9.1. *Среди всех многочленов $P_n(x)$, удовлетворяющих условию $\max\limits_{x\in[a,b]}|P_n(x)| = M$ приведенный многочлен Чебышёва $M\cdot T_n\left(\frac{2x-(a+b)}{b-a}\right)$ принимает максимальное значение для всех $\xi \notin (a,b)$, то есть*

$$|P_n(\xi)| < M\left|T_n\left(\frac{2\xi-(a+b)}{b-a}\right)\right|,\ \forall\ \xi\notin(a,b)$$

Доказательство. 1. Пусть $\exists\ \xi\notin(a,b),\ \exists\ P_n(x) : |P_n(\xi)| > M\left|T_n\left(\frac{2\xi-(a+b)}{b-a}\right)\right|$. Рассмотрим многочлен следующего вида

$$Q_n(x) = \underbrace{\frac{P_n(\xi)}{T_n\left(\frac{2\xi-(a+b)}{b-a}\right)}T_n\left(\frac{2x-(a+b)}{b-a}\right)}_{\tilde{T}_n} - P_n(x)$$

   (a) Многочлен $Q_n(x)\not\equiv 0$, так как $\|P_n(x)\| = M$, а $\left|\frac{P_n(\xi)}{T_n\left(\frac{2\xi-(a+b)}{b-a}\right)}\right| > M$.

   (b) Многочлен $Q_n(x)$ степени не больше n.

2. Обратим внимание, что $\operatorname{sgn} Q_n(x_{(m)}) = \operatorname{sgn}(\tilde{T}_n(x_{(m)})) = (-1)^m,\ m = 0,\dots,n$, где $x_{(m)} = \frac{a+b}{2} + \frac{b-a}{2}\cos\frac{\pi m}{n}$ - экстремумы приведенного многочлена Чебышёва. То есть Q_n имеет $n+1$ экстремум на $[a,b] \Rightarrow n$ корней на $[a,b]$.

3. Обратим так же внимание, что $Q_n(\xi) = 0$. То есть на отрезке $[a,b]$ Q_n имеет n корней и еще один корень в точке $\xi\notin(a,b) \Rightarrow Q_n(x)\equiv 0$. Противоречие.

□

Теорема 9.2 (Марков А.А)**.** *Среди всех многочленов $P_n(x)$, удовлетворяющих условию $\max\limits_{x\in[a,b]}|P_n(x)| = M$ производная приведенного многочлена Чебышёва $M\cdot T_n\left(\frac{2x-(a+b)}{b-a}\right)$ принимает максимальное значение для всех $\xi\notin(a,b)$, то есть*

$$|P_n'(x)|_{x=\xi} < M\left|T_n'\left(\frac{2x-(a+b)}{b-a}\right)\right|_{x=\xi}$$

Равенство достигается только для указанного полинома.

Доказательство. 1. Пусть $\exists\ \xi\notin(a,b),\ \exists\ P_n(x) : |P_n'(x)|_{x=\xi} > M\left|T_n'\left(\frac{2x-(a+b)}{b-a}\right)\right|_{x=\xi}$. Рассмотрим многочлен

$$Q_n(x) = \underbrace{\frac{P_n'(x)_{x=\xi}}{T_n'\left(\frac{2x-(a+b)}{b-a}\right)_{x=\xi}}T_n\left(\frac{2x-(a+b)}{b-a}\right)}_{\tilde{T}_n} - P_n(x)$$

   (a) Многочлен $Q_n(x)\not\equiv 0$, так как $\|P_n(x)\| = M$, а $\left|\frac{P_n'(x)_{x=\xi}}{T_n'\left(\frac{2x-(a+b)}{b-a}\right)_{x=\xi}}\right| > M$.

   (b) Многочлен $Q_n(x)$ степени не больше n.

2. Обратим внимание, что $\operatorname{sgn} Q_n(x_{(m)}) = \operatorname{sgn}(\tilde{T}_n(x_{(m)})) = (-1)^m,\ m = 0,\dots,n$, где $x_{(m)} = \frac{a+b}{2} + \frac{b-a}{2}\cos\frac{\pi m}{n}$ - экстремумы приведенного многочлена Чебышёва. То есть Q_n имеет $n+1$ экстремум на $[a,b] \Rightarrow n$ корней. Значит $Q_n'(x)$ имеет $n-1$ корень на $[a,b]$.

3. Обратим так же внимание, что

$$Q_n'(\xi) = \frac{P_n'(x)_{x=\xi}}{T_n'\left(\frac{2x-(a+b)}{b-a}\right)_{x=\xi}}T_n'\left(\frac{2x-(a+b)}{b-a}\right)_{x=\xi} - P_n'(x)_{x=\xi} = 0$$

   То есть на отрезке $[a,b]$ Q_n' имеет $n-1$ корней и еще один корень в точке $\xi\notin(a,b) \Rightarrow Q_n'(x)\equiv 0 \Rightarrow Q_n(x)\equiv \text{const}$ и Q_n имеет n корней $\Rightarrow Q_n\equiv 0$. Противоречие.

(Единственность экстремального полинома без доказательства). □

10 Конечно-разностный метод. Аппроксимация, устойчивость, сходимость, теорема Филиппова.

Пусть в области Ω с границей $\partial\Omega$ задана дифференциальная задача с граничным условием

$$\begin{cases} Ly = f, & x \in \Omega, \quad L : Y \to F,\ y \in Y,\ f \in F & (1) \\ ly = \varphi, & x \in \partial\Omega, \quad l : Y \to \Phi,\ \varphi \in \Phi & (2) \end{cases}$$

Здесь L и l - дифференциальные операторы; f и φ заданные элементы, а y – искомый элемент некоторых линейных нормированных пространств F, Φ и Y с заданными нормами $\|\cdot\|_F$, $\|\cdot\|_\Phi$ и $\|\cdot\|_Y$ соответственно.

Конечно-разностный метод.

Для применения разностного метода задают некоторую *сетку* - конечное множество точек (узлов) $\overline{\Omega}_h = \Omega_h \cup \partial\Omega_h$, принадлежащее области $\overline{\Omega} = \Omega \cup \partial\Omega$, определяют сеточные пространства F_h, Φ_h и Y_h и задают операторы проектирования $(\cdot)_{Y_h} : Y \to Y_h, (\cdot)_{F_h} : F \to F_h, (\cdot)_{\Phi_h} : \Phi \to \Phi_h$ элементов исходных пространств на элементы сеточных пространств.

• Далее при написании операторов проектирования будем опускать на какое пространство он действует, так как из контекста будет понятно: если пишем $(f)_h$, то имеем в виду $(f)_{F_h}$.

При этом в пространствах задаются согласованные нормы.

Опр. 10.1. Нормы $\|\cdot\|$, U и $\|\cdot\|_h$, U_h называются согласованными, если для произвольной достаточно гладкой функции u выполняется соотношение

$$\lim_{h\to 0} \|(u)_h\|_h = \|u\|$$

Пример 10.1. Рассмотрим $y(x) \in C[0,1]$, выберем норму $\|y\|_C = \max\limits_{x\in[0,1]} |y(x)|$.

Зададим оператор проектирования $(y)_h \to [y(0), y(h), \ldots, y(Nh)] \in Y_h$. Для векторов в $\mathbb{R}^{n+1}$ возьмем евклидову норму: $\|y_h\|_{\mathbb{R}^{n+1}} = \sqrt{\sum\limits_{i=0}^{N} y_i^2}$.

Проверим согласованность: рассмотрим непрерывную функцию $y^0 \equiv 1$:

$$\begin{array}{rcccl} \|y^0\|_C & = & \max\limits_{x\in[0,1]} |1| & = & 1 \\ \|(y^0)_h\|_{\mathbb{R}^{N+1}} & = & \sqrt{\sum\limits_{i=0}^{N} 1} & = & \sqrt{N+1} \end{array} \quad \Rightarrow \lim_{h\to 0} \|(y)_h\|_h = \infty \neq \|(y^0)\|_C = 1$$

То есть норма $\|y_h\|_{\mathbb{R}^{n+1}} = \sqrt{\sum\limits_{i=0}^{N} y_i^2}$ не является согласованной с $\|y\|_C = \max\limits_{x\in[0,1]} |y(x)|$.

Для нормы $\|y\|_C = \max\limits_{x\in[0,1]} |y(x)|$ в качестве согласованной часто выбирают $\|y_h\|_{\mathbb{R}^{n+1}} = \max\limits_i |y_i|$

Замечание. Если нормы не являются согласованными, то из условия $\lim\limits_{h\to 0} \|(u)_h\|_h = 0$ не следует $\|u\| = 0$ т.е. что $u \equiv 0$ в исходном пространстве U. Однако, это требование необходимо для обоснования сходимости сеточной функций u_h к непрерывной u.

• Далее когда будем говорить о нормах $\|\cdot\|_{F_h}$, $\|\cdot\|_{Y_h}$, $\|\cdot\|_{\Phi_h}$, будем иметь в виду, что они согласованы соответственно с $\|\cdot\|_F$, $\|\cdot\|_Y$, $\|\cdot\|_\Phi$.

Все производные, входящие в уравнение и краевые условия, заменяются *разностными аппроксимациями*. В результате дифференциальные операторы L и l заменяются разностными L_h и l_h. Для нахождения приближенного решения задачи (1), (2) определим *разностную схему* - семейство разностных задач, зависящих от параметра h:

$$\begin{cases} L_h y_h = f_h, & x_h \in \Omega_h, \quad L_h : Y_h \to F_h,\ y_h \in Y_h,\ f_h \in F_h & (3) \\ l_h y_h = \varphi_h, & x_h \in \partial\Omega_h, \quad l_h : Y_h \to \Phi_h,\ \varphi_h \in \Phi_h & (4) \end{cases}$$

Пример 10.2. Рассмотрим в качестве примера дифференциальную задачу, $\|y(x)\|_{C^\infty} = \max\limits_{x\in\mathbb{R}} |y(x)|$

$$\begin{cases} y'(x) = e^x, & x \in \Omega \equiv [0,1], \quad L \equiv \frac{\partial}{\partial x}, \quad y \in Y \equiv C^\infty,\ f \in F \equiv C^\infty & (1) \\ y(0) = 1, & x \in \partial\Omega \equiv \{0,1\}, \quad l \equiv I, \quad \varphi \in \Phi \equiv C^\infty & (2) \end{cases}$$

Предлагается для численного решения взять шаг сетки $h = \frac{1}{N}$, операторы проектирования:

$$\begin{array}{lcc} (y)_{Y_h} & = & [y(x_0), \dots, y(x_N)] \\ (f)_{F_h} & = & [\frac{f(x_1)-f(x_0)}{2}, \dots, \frac{f(x_N)-f(x_{N-1})}{2}] \\ (\varphi)_{\Phi_h} & = & \varphi(x_0) \end{array}$$

в качестве согласованных норм выбрать $\|y_h\|_{Y_h} = \max\limits_i |y_i|$, $Y_h \equiv \mathbb{R}^{N+1}$, $\|f_h\|_{F_h} = \max\limits_i |f_i|$, $F_h \equiv \mathbb{R}^N$ и рассмотреть разностную схему

$$\begin{cases} \frac{y_{k+1}-y_k}{h} = \frac{e^{x_k+1}-e^{x_k}}{2} =: f_k \quad x_k \in \Omega_h \equiv \{0, h, \dots h(N-1)\}, \quad (L_h y_h)_i \equiv \frac{y_{i+1}-y_i}{h} & (3) \\ y_0 = 1, \qquad\qquad x_k \in \partial\Omega_h \equiv \{0, N-1\}, \qquad\qquad l_h \equiv I, & (4) \end{cases}$$

Задача (3), (4) образует систему линейных уравнений, у которой $\exists!$ решение y_k, но как $y_k \in Y_h$ хоть как-то связано с $y \in Y$? Почему задачи (1), (2) хоть как-то связаны с задачами (3), (4)?

Сходимость

Опр. 10.2. Решение y_h разностной задачи (3), (4) сходится к решению y дифференциальной задачи (1), (2), если $\exists\ h_0, c, p$ такие что

$$\|(y)_{Y_h} - y_h\|_{Y_h} \le ch^p,\ \forall\ h \le h_0$$

причем c и p не зависят от h. Число p называют порядком сходимости разностной схемы.

Доказывать сходимость в общем случае может быть не просто, ее удобно свести к проверке аппроксимации и устойчивости, а затем воспользовать теоремой Филиппова.

Аппроксимация

Опр. 10.3. Разностная задача (3), (4) аппроксимирует дифференциальную задачу (1), (2), с порядком аппроксимации $p = \max(p_1, p_2)$, если $\exists\ h_0, c_1, c_2, p_1, p_2$:

$$\begin{array}{l} \|L_h(y)_{Y_h} - (Ly)_{F_h}\|_{F_h} + \|(f)_{F_h} - f_h\|_{F_h} \le c_1 h^{p_1} \\ \|l_h(y)_{Y_h} - (ly)_{\Phi_h}\|_{\Phi_h} + \|(\varphi)_{\Phi_h} - \varphi_h\|_{\Phi_h} \le c_2 h^{p_2} \end{array} \ \forall\ h \le h_0$$

При этом c_1, c_2, p_1, p_2 не зависят от h.

Опр. 10.4. Разностный оператор L_h из (3) локально аппроксимирует дифференциальный оператор L из (1) в точке x_i, если для достаточно гладкой функции y $\exists\ h_0, c, p$:

$$|(L_h(y)_{Y_h} - (Ly)_{F_h})_{x=x_i}| \le ch^p,\ \forall\ h \le h_0$$

Опр. 10.5. Говорят, что разностная схема (3), (4) аппроксимирует на решении дифференициальную задачу (1), (2) с порядком аппроксимации $p = \min(p_1, p_2)$, если $\exists\ h_0, c_1, c_2, p_1, p_2$:

$$\begin{array}{l} \|L_h(y)_{Y_h} - f_h\|_{F_h} \le c_1 h^{p_1} \\ \|l_h(y)_{Y_h} - \varphi_h\|_{\Phi_h} \le c_2 h^{p_2} \end{array} \ \forall\ h \le h_0$$

При этом c_1, c_2, p_1, p_2 не зависят от h и выполнено условие нормировки

$$\|(f)_{F_h} - f_h\|_{F_h} \underset{h\to 0}{\to} 0$$

$$\|(\varphi)_{\Phi_h} - \varphi_h\|_{\Phi_h} \underset{h\to 0}{\to} 0$$

Лемма 10.1. *Аппроксимация задач $\Rightarrow$ аппроксимация на решении.*

Доказательство. Записываем определение аппроксимации на решении, получаем оценку сверху.

$$\begin{array}{ccccccccc} \|L_h(y)_h - f_h\|_h & = & \|L_h(y)_h - (Ly)_h + (Ly)_h - f_h\|_h & \le & \|L_h(y)_h - (Ly)_h\|_h & + & \|(Ly)_h - f_h\|_h & = \\ \|l_h(y)_h - \varphi_h\|_h & & \|l_h(y)_h - (ly)_h + (ly)_h - \varphi_h\|_h & & \|l_h(y)_h - (ly)_h\|_h & + & \|(ly)_h - \varphi_h\|_h & \end{array}$$

$$\begin{array}{ccccc} & & \|L_h(y)_h - (Ly)_h\|_h & + & \|(f)_h - f_h\|_h \le c_1 h^{p_1} \\ & = & \|l_h(y)_h - (ly)_h\|_h & + & \underset{\text{условия нормировки}}{\|(\varphi)_h - \varphi_h\|_h} \le c_2 h^{p_2} \end{array} \ \forall\ h \le h_0 \text{ - аппроксимация задач}$$

□

Устойчивость

Опр. 10.6. Разностная схема (3), (4) называется устойчивой если $\forall\ \varepsilon > 0$ можно подобрать $\delta = \delta(\varepsilon)$ такое, что для произвольных решений $y_h^{(1)}$, $y_h^{(2)}$ при $\forall\ h \le h_0$

$$\left\|f_h^{(1)} - f_h^{(2)}\right\|_{F_h} + \left\|\varphi_h^{(1)} - \varphi_h^{(2)}\right\|_{\Phi_h} \le \delta \Rightarrow \left\|y_h^{(1)} - y_h^{(2)}\right\|_{Y_h} \le \varepsilon$$

То есть малое изменение во входных данных не влечет большого изменения в решении.

Опр. 10.7. Линейная схема (3), (4) называется устойчивой если для произвольных решений $y_h^{(1)}$, $y_h^{(2)}$ при $\forall\ h \le h_0\ \exists c_1, c_2$

$$\left\|y_h^{(1)} - y_h^{(2)}\right\|_{Y_h} \le c_1 \left\|f_h^{(1)} - f_h^{(2)}\right\|_{F_h} + c_2 \left\|\varphi_h^{(1)} - \varphi_h^{(2)}\right\|_{\Phi_h}$$

Замечание 10.1. Дали задачу $\begin{cases} L_h y_h = f_h \\ l_h y_h = \varphi_h \end{cases}$. Если L_h, l_h - линейные, то можем перейти к системе линейных уравнений $A_h y_h = b_h,\ A_h \equiv \begin{pmatrix} L_h \\ l_h \end{pmatrix},\ b_h \equiv \begin{pmatrix} f_h \\ \varphi_h \end{pmatrix}$. Тогда если задача устойчива, то константу можно искать следующим образом:

$$A_h(y_h^{(1)} - y_h^{(2)}) = b_h^{(1)} - b_h^{(2)} \Rightarrow y_h^{(1)} - y_h^{(2)} = A_h^{-1}(b_h^{(1)} - b_h^{(2)}) \Rightarrow \left\|y_h^{(1)} - y_h^{(2)}\right\|_{Y_h} \le \left\|A_h^{-1}\right\|_h \left\|b_h^{(1)} - b_h^{(2)}\right\|_h$$

То есть устойчивость для линейных разностных схем обозначает, что $\left\|A_h^{-1}\right\|_h \le \text{const} \le \infty,\ h \to 0$.

Пример. Рассмотрим дифференциальную задачу $y(x) = f(y)$. Построим для нее разностную схему $h \cdot y_h = h \cdot f_h$.

- Тогда $(L_h y_h)_i \equiv h y_i$, но $\|L_h\|^{-1} = \frac{1}{h} \underset{h\to 0}{\to} \infty$, то есть данная разностная схема не устойчива.
- Обратим так же внимание, что $\|(f)_{F_h} - f_h\|_{F_h} = \|h \cdot f_h - f_h\|_{F_h} \underset{h\to 0}{\not\to} 0$, то есть мы так же не можем ничего сказать про аппроксимацию задач, так как не соблюдена нормировка правых частей.

Замечание 10.2. Дана задача $Ax = b$, тогда по теореме Кронекера-Капелли: если при $b \equiv 0 \Rightarrow \exists!\ x:\ Ax = b,\ x \equiv 0$, то $\forall\ b\ \exists!\ x:\ Ax = b$.

Рассмотрим два решения $y_h^{(1)} \equiv y_h$ и $y_h^{(2)} \equiv 0$: $\begin{cases} A_h y_h = b_h \\ A_h \cdot 0 = 0 \end{cases} \underset{\text{устойчивость}}{\Rightarrow} \|y_h\|_h \le C \|b_h\|_h$. Тогда по теореме Кронекера-Капелли если из $b_h \equiv 0 \Rightarrow y_h \equiv 0 \Rightarrow \forall\ b_h \exists!\ y_h$. То есть таким образом **устойчивость задачи влечет ее корректность**.

Теорема 10.1 (Филиппов А.Ф.)**.** *Пусть выполнены следующие условия*

1. *L, l, L_h, l_h - линейные операторы.*
2. *Существует и единственно решение дифференциальной задачи* (1), (2).
3. *Разностная схема* (3), (4) *аппроксимирует задачу* (1), (2) *на решении с порядком p.*
4. *Разностная схема* (3), (4) *устойчива*

Тогда решение разностной задачи (3), (4) *сходится к решению* (1), (2) *с порядком не ниже p.*

Доказательство. Запишем две задачи

$$\begin{cases} L_h y_h = f_h \\ l_h y_h = \varphi_h \end{cases} \qquad \begin{cases} L_h(y)_h = f_h + (L_h(y)_h - f_h) \\ l_h(y)_h = \varphi_h + (l_h(y)_h - \varphi_h) \end{cases}$$

- Так как разностная схема устойчива, то решение $y_h\ \exists!$
- $(y)_h$ валидная запись, так как решение дифференциальной задачи $\exists!$ по условию.

Так как операторы L, l, L_h, l_h линейные, то запишем оценку из определения устойчивости:

$$\begin{aligned} \|y_h - (y)_h\|_{Y_h} &\le c_1 \|f_h - (f_h - (L_h(y)_h - f_h))\|_{F_h} + c_2 \|\varphi_h - (\varphi_h - (l_h(y)_h - \varphi_h))\|_{\Phi_h} = \\ &= c_1 \underbrace{\|L_h(y)_h - f_h\|_{F_h}}_{\le \hat{c} h^p} + c_2 \underbrace{\|l_h(y)_h - \varphi_h\|_{\Phi_h}}_{\le \hat{c} h^p} \underset{\text{аппроксимация}}{\le} \hat{c}(c_1 + c_2)h^p \le \overline{c} h^p \end{aligned}$$

□

11 Метод неопределенных коэффициентов построения разностных схем. Погрешность формул численного дифференцирования, оценка для оптимального шага.

Метод неопределенных коэффициентов построения разностных схем

Дали задачу $Ly = f,\ L \equiv \frac{d^m}{dx^m}$. Надо перейти к разностной схеме $L_h y_h = f_h$ в точках $\{x_i\}$, $f_h = (f)_{F_h}$.

Оператор $L_h y_h$ будем строить в виде $\frac{1}{h^m}\sum\limits_{i=0}^{N} c_i y_k$ на шаблоне x_i, а коэффициенты будем искать из условия аппроксимации на задаче $|(Ly)_{F_h} - L_h(y)_{Y_h}| + |f_h - \underset{=0}{(f)_{F_h}}| \le ch^p$ с желанием получить наивысший p, то есть

$$\left| y^{(m)}(x_i) - \frac{1}{h^m}\sum_{i=0}^{N} c_i y(x_i) \right| \le ch^p$$

Представим $x_i = x_k + \alpha_i h$ и разложим каждое $y(x_i)$ в ряд Тейлора в узле сетки x_k в виде $y(x_k + \alpha_i h) = y(x_k) + y'(x_k)\alpha_i h + \ldots + \underline{\underline{\mathcal{O}}}(h^{p+m})$. Подставив данные разложения в условия аппроксимации получим систему линейных уравнений на c_i.

Пример 11.1. Найти коэффициенты разностного оператора при наибольшем p

$$y'(x_k) = \frac{1}{h}(c_1 y(x_k - h) + c_2 y(x_k) + c_3 y(x_k + h)) + \underline{\underline{\mathcal{O}}}(h^p)$$

Разложим в ряд Тейлора с остаточным членом в форме Лагранжа в точке x_k:

$$y(x_k \pm h) = y(x_k) \pm y'(x_k)h + y''(x_k)\frac{h^2}{2} \pm y'''(\xi_\pm)\frac{h^3}{6}$$

Подставим в исходное уравнение и сопоставим коэффициенты при соответствующих слагаемых слева и справа

$$\begin{cases} y(x_k): & c_1 + c_2 + c_3 = 0 \\ y'(x_k): & c_3 - c_1 = 1 \\ y''(x_k): & c_1 + c_3 = 0 \\ y'''(x_k): & c_3 - c_1 = 0\ \times \end{cases} \Rightarrow (\star)$$

Обращаем внимание, что последнее уравнение противоречит второму, то есть мы сможем наиграть с помощью коэффициентов только второй порядок.

$$(\star) \Rightarrow \begin{cases} c_1 = -\frac{1}{2} \\ c_2 = 0 \\ c_3 = \frac{1}{2} \end{cases}$$

Итоговая разностная схема с порядком аппроксимации $p = 2$ имеет вид

$$L_h y_h = \frac{y_{k+1} - y_{k-1}}{2h}$$

Получим оценку для константы:

$$\left| \frac{1}{h}(c_3 y'''(\xi_+)\frac{h^3}{6} - c_1 y'''(\xi_-)\frac{h^3}{6}) \right| = \frac{h^2}{12}\left| y'''(\xi_+) + y'''(\xi_-) \right| \le ch^2 \Rightarrow c \ge \frac{|y'''(\xi_+) + y'''(\xi_-)|}{12}$$

Отсюда в том числе следует, что найденная схема точна для произвольного многочлена второй степени. Таким образом, систему уравнений на коэффициенты можно найти из условия точности формулы разностного дифференцирования для многочленов наиболее высокой степени. Для этого подставляем последовательно $y(x) = 1, x, x^2, \ldots$ в разностную формулу и приравниваем к точному значению производной $y^{(m)}(x)$. Решение полученной линейной системы определяет те же коэффициенты схемы.

Пример 11.2. Найти коэффициенты разностного оператора при наибольшем p

$$y'(x_k) = \frac{1}{h}(c_1 y(x_k - h) + c_2 y(x_k) + c_3 y(x_k + h)) + \underline{\underline{\mathcal{O}}}(h^p)$$

$$\begin{cases} 1: & c_1 + c_2 + c_3 = 0 \\ x: & c_1(x_k - h) + c_2(x_k) + c_3(x_k + h) = h \\ x^2: & c_1(x_k - h)^2 + c_2(x_k)^2 + c_3(x_k + h)^2 = 2x_k h \\ x^3: & c_3 - c_1 = 0\ \times \end{cases} \Rightarrow (\star)$$

Решая данную систему должны получить те же самые коэффициенты.

Погрешность формул численного дифференцирования, оценка для оптимального шага.

Взяли в качестве аппроксимации для $y^{(m)}$ следующее выражение

$$y^{(m)}(x_k) = \underbrace{\frac{1}{h^m}\sum_{i=0}^{N} c_i y(x_k + \alpha_i h)}_{D} + \underbrace{ch^p}_{E_1(h)}$$

Вносим значения в компьютер, в котором имеем погрешность ε_i: $y(x_k) \to y(x_k) + \varepsilon_i$.

Получаем следующую величину:

$$y^{(m)}(x_k) = \frac{1}{h^m}\sum_{i=0}^{N} c_i(y(x_k + \alpha_i h) + \varepsilon_i) + E_1 = D + \underbrace{\frac{1}{h^m}\sum_{i=0}^{N} c_i\varepsilon_i}_{E_2(h)} + E_1(h)$$

Таким образом итоговая погрешность на компьютере равна $E_2 + E_1$. Так как ε_i - машинная точность - по своей сути случайна, то точно оценить мы не можем, сделаем это грубо

$$|E_1(h) + E_2(h)| \le \frac{\max_i |\varepsilon_i|}{h^m}\sum_{i=0}^{N}|c_i| + ch^p \le \frac{A\varepsilon}{h^m} + ch^p =: E(h),\ \varepsilon = \max_i |\varepsilon_i|,\ \sum_{i=0}^{N}|c_i| \le A$$

Обратим внимание, что при достаточно больших h второе слагаемое будет вносить большой вклад в погрешность, а при малых h - первое. Хотим найти оптимальное значение h_0:

$$E'(h_0) = 0 = \frac{-mA\varepsilon}{h_0^{m+1}} + cph_0^{p-1} \Leftrightarrow \frac{mA\varepsilon}{h_0^{m+1}} = cph^{p-1} \Rightarrow h_0 = \left(\frac{mA\varepsilon}{cp}\right)^{\frac{1}{p+m}} \sim \varepsilon^{\frac{1}{p+m}}$$

$$E(h_0) \sim \varepsilon^{1-\frac{m}{p+m}} + \varepsilon^{\frac{p}{p+m}} \sim \varepsilon^{\frac{p}{p+m}}$$

Пример 11.3. Известно, что

$$y''(x_k) = \frac{y(x_k + h) - 2y(x_k) + y(x_k - h)}{h^2} + ch^2$$

Пусть погрешность вычислений ϵ не превышает 10^{-4}. В наших терминах $p = 2$, $m = 2$. Оптимальный шаг $h_0 = (10^{-4})^{1/4} = 10^{-1}$ будет влечь погрешность $E = 10^{-2}$.

12 Задача Коши, условия аппроксимации p-го порядка на решении, α-устойчивость. Модельные схемы.

Опр. 12.1. Задачей Коши первого порядка называют следующую дифференциальную задачу

$$\begin{cases} y'(x) = f(y(x), x),\ y \in C^{(m)}[x_0, x_0 + X],\ \|y\| = \max\limits_{x \in [x_0, x_0+X]} |y(x)| \\ y(x_0) = y^0 \end{cases}$$

Для такой задачи предлагается строить следующую разностную схему

$$h = \frac{X}{N},\ x_i = x_o + i \cdot h,\ y_h = \{y_i\}_{i=0}^N,\ \|y_h\|_{Y_h} = \max_i |y_k|$$

$$\begin{cases} \frac{1}{h} \sum\limits_{i=0}^{n} a_{-i} y_{k-i} = \sum\limits_{i=0}^{n} b_{-i} f(y_{k-i}, x_{k-i}) \\ y_0, y_1, \ldots, y_{n-1} - \text{ начальные условия (может быть больше одного)} \end{cases} \tag{1}$$

Если $a_0 \neq 0,\ b_0 \neq 0$, то схема называется неявной, если $a_0 \neq 0,\ b_0 = 0$, то явной, а если $a_0 = 0,\ b_0 \neq 0$ с забеганием вперед.

Хотим для такой разностной схемы проверять условие аппроксимации на решении, чтобы далее пользоваться теоремой Филиппова.

Зададим функцию погрешности

$$r_h^k \overset{\text{def}}{=} \frac{1}{h} \sum_{i=0}^{n} a_{-i} y(x_{k-i}) - \sum_{i=0}^{n} b_{-i} f(y(x_{k-i}), x_{k-i})$$

Условия аппроксимации на решении с порядком p на отрезке $[x_0, x_0 + X]$: $\begin{cases} \|r_h\|_{F_h} \leq ch^p \\ \|f_h - (f)_h\|_{F_h} \underset{h \to 0}{\to} 0 \end{cases}$

Коэффициенты в общем случае a_{-i} и b_{-i} находятся из условий аппроксимации на задаче

$$\|L_h(y)_{Y_h} - (Ly)_{F_h}\|_{F_h} + \|(f)_{F_h} - f_h\|_{F_h} \leq c_1 h^{p_1}$$

.

Замечание 12.1. Так как задача Коши по сути сводится к интегрированию f, то рассуждения о погрешности формул численного дифференицирования здесь неприменимы.

Теорема 12.1 (Условия аппроксимации p-го порядка на решении)**.** *Для задачи $y'(x) = f(x)$ с разностной схемой* (1) *условия аппроксимации p-го порядка на решении имеют вид*

$$\sum_{i=0}^{n} a_{-i} = 0;\ \sum_{i=0}^{n} b_{-i} = 1;\ \sum_{i=0}^{n} a_{-i} i = -1;$$

$$\sum_{i=0}^{n} (a_{-i} i + b_{-i} s) i^{s-1} = 0,\ s = 2, \ldots, p$$

Доказательство. 1. Проверим условие нормировки правых частей

$$\|f_h - (f)_h\|_{F_h} = \left\| f(x_{k-i}) - \sum_{i=0}^{n} b_{-i} f(x_{k-i}) \right\|_{F_h} = \left\| f(x_{k-i}) \left(1 - \sum_{i=0}^{n} b_{-i} \right) \right\|_{F_h} \underset{h \to 0}{\to} 0 \Leftrightarrow \sum_{i=0}^{n} b_{-i} = 1$$

2. Выпишем ряд Тейлора для левой и правой части в узлах $\{x_k\}$:

$$y(x_k - ih) = y(x_k) - ihy'(x_k) + \sum_{s=2}^{p} \frac{(-ih)^s y^{(s)}(x_k)}{s!} + \underline{\underline{\mathcal{O}}}(h^{p+1})$$

$$f(x_k - ih) = y'(x_k - ih) = y'(x_k) + \sum_{s=2}^{p} \frac{(-ih)^{s-1} y^{(s)}(x_k)}{(s-1)!} + \underline{\underline{\mathcal{O}}}(h^p)$$

Запишем условие аппроксимации на решении до p-го порядка:

$$\left|\frac{1}{h}\sum_{i=0}^{n} a_{-i}y(x_k - ih) - \sum_{i=0}^{n} b_{-i}f(x_k - ih)\right| = \left|\frac{1}{h}\sum_{i=0}^{n} a_{-i}y(x_k) +\right.$$

$$\left. +\frac{1}{h}\sum_{i=0}^{n} a_{-i}(-ihy'(x_k)) - \sum_{i=0}^{n} b_{-i}y'(x_k) + \frac{1}{h}\sum_{i=0}^{n} a_{-i}\sum_{s=2}^{p}\frac{(-ih)^s y^{(s)}(x_k)}{s!} - \sum_{i=0}^{n} b_{-i}\sum_{s=2}^{p}\frac{(-ih)^{s-1}y^{(s)}(x_k)}{(s-1)!}\right| =$$

$$= \left|\frac{y(x_k)}{h}\sum_{i=0}^{n} a_{-i} + y'(x_k)\left(-\sum_{i=0}^{n} a_{-i}i - \sum_{i=0}^{n} b_{-i}\right) + \sum_{s=2}^{p}\frac{(-h)^{s-1}y^{(s)}}{s!}\left(\sum_{i=0}^{n}(a_{-i}i + b_{-i}s)i^{s-1}\right)\right| \le ch^p$$

Для выполнения условия аппроксимации нужно, чтобы все коэффициенты до h^p занулились, отсюда и из проверки нормировки следует доказательство теоремы.

□

Замечание 12.2. Для уверенности, что схема аппроксимирует (т.е. $p = 1$) необходимо и достаточно проверить первые три уравнения.

Для того, чтобы найти a_{-i} и b_{-i} надо решить систему уравнений. Для того, чтобы система была разрешима, важно проверить $p = 2n$.

Пример 12.1. Схемы Адамса со вторым порядком аппроксимации на решении:

$$\begin{array}{rrl} \text{явная схема}: & \frac{y_k - y_{k-1}}{h} = & \frac{3}{2}f_{k-1} - \frac{1}{2}f_{k-2} \\ \text{неявня схема}: & \frac{y_k - y_{k-1}}{h} = & \frac{1}{2}\left(f_k + f_{k-1}\right) \end{array}$$

Устойчивость разностных схем для задач Коши довольно затруднительно проверить по определению, используют более слабое понятие α-устойчивости.

Опр. 12.2. Разностная схема $\frac{1}{h}\sum\limits_{i=0}^{n} a_{-i}y_{k-i} = f_k$ для задачи $y'(x) = f(x)$ называется α-устойчивой, если все корни соответствующего характеристического многочлена однородного разностного уравнения принадлежат единичному кругу и на границе нет кратных корней.

Замечание 12.3. α-устойчивость строго говоря не связана с устойчивостью, но для широкого класса задач обеспечивает сходимость.

Пример 12.2. Рассмотрим схемы с α-устойчивостью, но с разными результатами. Будем численно решать задачу Коши $y' = -y,\ y(0) = 1,\ x = [0, X]$ с точным решением $y(x) = e^{-x}$. Проверим разные разностные схемы и оценим их качество.

1. Неявная схема Адамса: $n = 1,\ p = 2$

$$\begin{cases} \frac{y_k - y_{k-1}}{h} = \frac{1}{2}(f_k + f_{k-1}) \\ y_0 = y^0 \end{cases} \Leftrightarrow \begin{array}{c} x_k = x_0 + ih,\ h = \frac{X}{n} \\ ((y)_{Y_n})_k = y(x_k) =: y_k \\ f(y_k, x_k) = -y_k =: f_k \\ ((f)_{F_n})_k = \frac{1}{2}(f_k + f_{k-1}) \end{array} \Leftrightarrow \begin{cases} \frac{y_k - y_{k-1}}{h} = -\frac{y_k + y_{k-1}}{2} \\ y_0 = 1 \end{cases}$$

Проверим условие аппроксимации второго порядка:

$$\sum_{i=0}^{n} a_{-i} = 1 - 1 = 0;\ \sum_{i=0}^{n} b_{-i} = \frac{1+1}{2} = 1;\ \sum_{i=0}^{n} a_{-i} i = 1 \cdot 0 - 1 \cdot 1 = -1$$
$$\sum_{i=0}^{n} (a_{-i} i + b_{-i} \cdot 2) i^{2-1} = -1 + \frac{1}{2} \cdot 2 = 0$$

Проверим α-устойчивость: $y_k - y_{k-1} = 0 \Rightarrow P(\mu) = \mu - 1 = 0 \Rightarrow \mu = 1$

Таким образом разностная схема должна сходиться. Мы можем это проверить, решив эквивалентное однородное разностное уравнение с постоянными коэффициентами:

$$\frac{y_k - y_{k-1}}{h} = -\frac{y_k + y_{k-1}}{2} \Leftrightarrow y_k(2 - h) + y_{k-1}(2 + h) = 0;\ P(\mu) = 0 \Rightarrow \mu = \frac{1 - h/2}{1 + h/2} \Rightarrow y_k = \left(\frac{1 - h/2}{1 + h/2}\right)^k$$

$$y_n = \left(\frac{1 - h/2}{1 + h/2}\right)^{X/h} \underset{h \to 0}{\overset{\text{Тейлор в 0}}{\longrightarrow}} \left(1 - h + \underline{\underline{\mathcal{O}}}(h^2)\right)^{X/h} \underset{\text{2й замечательный}}{\longrightarrow} e^{-X} = y(x_n)$$

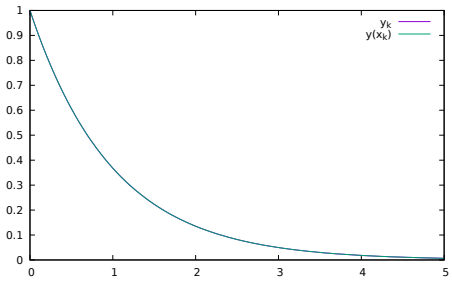


Рис. 4: Численные значения при $X = 5$, $n = 100$

2. Вторая схема аппроксимации

$$\begin{array}{c} ((y)_{Y_n})_k = y(x_k) =: y_k \\ ((f)_{F_n})_k = f_k := f(x_k) = -y(x_k) \end{array} \Rightarrow \begin{cases} \frac{y_{k+1} - y_{k-1}}{2h} + y_k = 0 \\ y_0 = 1,\ y_1 = 1 - h \end{cases}$$

(a) Аппроксимацию второго порядка на решении можно проверить, используя теорему выше, но для разнообразия проверим по определению

$$\|L_h(y)_{Y_h} - f_h\|_{F_h} = \max_k \left| \frac{y(x_k + h) - y(x_k - h)}{2h} - f(x_k) \right| =$$
$$\left| \begin{array}{c} y(x_k \pm h) = y(x_k) \pm y'(x_k)h + y''(x_k)\frac{h^2}{2} \pm y'''(x_k)\frac{h^3}{6} + \underline{\underline{\mathcal{O}}}(h^4) \\ f(x_k) = y'(x_k) - \text{по условию} \end{array} \right|$$
$$= \max_k \left| \frac{2hy'(x_k) + \underline{\underline{\mathcal{O}}}(h^3)}{2h} - y'(x_k) \right| \le ch^2$$

(b) Проверим аппроксимацию на краях $\|l_h(y)_h - \varphi_h\|_{\Phi_h}$

$$|y(0) - y_0| = 0 \le ch^2$$
$$|y(h) - y_1| = |\overbrace{y(0)}^{=1} + y'(0)h + \underline{\underline{\mathcal{O}}}(h^2) - 1 + h| = |h\overbrace{(y'(0) + y(0))}^{=0} + \underline{\underline{\mathcal{O}}}(h^2)| =\le ch^2$$

(c) Проверим условие нормировки: $\|(f)_h - f_h\|_{F_h} = |f(x_k) - f(x)| \to 0$

Разностная схема имеет второй порядок аппроксимации.

$$\frac{y_{k+1} - y_{k-1}}{2h} = 0 \Rightarrow P(\mu) = \mu^2 - 1 \Rightarrow \mu = \pm 1$$

Схема α-устойчива.

Судя по всей той теории, что описана выше, так как есть α-устойчивость, то такую разностную схему можно применять для численного подсчета. Давайте попробуем посчитать иначе: можем написать решение разностного уравнения второго порядка с постоянными коэффициентами.

$$\frac{y_{k+1} - y_{k-1}}{2h} + y_k = 0 \Rightarrow P(\mu) = \frac{\mu^2 - 1}{2h} + \mu = 0 \Rightarrow \mu_{1,2} = -h \pm \sqrt{1 + h^2}$$

$$y_k = C_1(-h + \sqrt{1 + h^2})^k + C_2(\underbrace{-h - \sqrt{1 + h^2}}_{<-1})^k$$

Обратим внимание, что из-за того, что второй корень по модулю больше единицы, при большом количестве шагов правое слагаемое в сумме будет доминировать, в результате чего решение разностной задачи будет болтаться в зависимости от четности k. То есть эту схему лучше не применять.

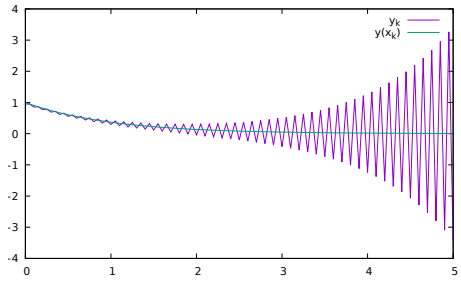


Рис. 5: Численные значения при $X = 5$, $n = 100$

3. Третья схема

$$\begin{cases} \frac{y_{k+1} - y_{k-1}}{2h} = -\frac{y_{k+1} + y_{k-1}}{2} \\ y_0 = 1, \ y_1 = 1 - h \end{cases}$$

Аналогично предыщум пунктам проверяется аппроксимация на решении 2 степени и α-устойчивость. Решения разностного уравнения имеют вид

$$\mu_{1,2} = \pm\sqrt{\frac{1 - h}{1 + h}}, \ |\mu_{1,2}| < 1$$

Заметим, что в данной схеме каждый следующий элемент порождается через предпоследний, из-за чего решение задачи будет менять знак в зависимости от k. Такая схема пригодна для расчетов, но в отличие от первой для нее нужно дополнительное краевое условие.

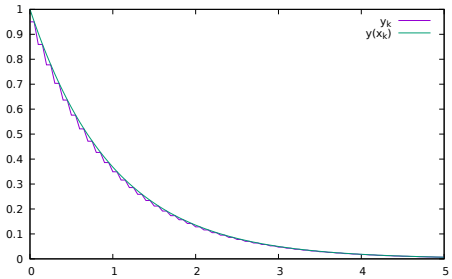


Рис. 6: Численные значения при $X = 5$, $n = 100$

13 Численные методы решения задачи Коши: метод Тейлора, методы Адамса.

Метод Тейлора

Решаем задачу Коши $\begin{cases} y'(x) = f(x,y) \\ y(0) = y^0 \end{cases}$ Мы хотим построить наше решение в точках $y(h), y(2h)\ldots$, предлагается явно посчитать значения в этих точках, разложив в ряд Тейлора до $p+1$ степени. Далее для упрощения записи $y \stackrel{\text{def}}{=} y(x_0),\ f \stackrel{\text{def}}{=} f(x_0, y(x_0))$:

$$\begin{cases} y(x) = y + hy' + \frac{h^2}{2}y'' + \sum_{s=2}^{p} y^{(s)}\frac{h^s}{s!} + \underline{\underline{\mathcal{O}}}(h^{p+1}) \\ y' = f \text{ – по условию} \\ y'' = f_x + f_y y' \\ y''' = f_{xx} + f_{xy}y' + f_{yx}y' + f_y y'' + f_{yy}(y')^2 \\ \ldots \end{cases}$$

Метод применим если $|x - x_0|$ больше области сходимости ряда Тейлора, иначе предлагается разбивать отрезки на подотрезки и строить решение в точке x за несколько шагов.

Замечание 13.1. Данный алгоритм может быть полезен, когда требуется решить большое количество задач вполне определенного вида с различными начальными данными. В этом случае требуемые производные можно найти аналитически и сохранить для многократного применения.

Пример 13.1. Решаем задачу $y'(x) = x^2\sin(y(x))$. Мы можем посчитать

$$\begin{aligned} y''(x) &= 2x\sin(y(x)) + x^2\cos(y(x))y'(x) \\ &= 2x\sin(y(x)) + x^4\cos(y(x))\sin(y(x)) \end{aligned}$$

Таким образом второй порядок *локальной* точности используя следующую формулу

$$y_{k+1} = y_k + hx_k^2\sin(y_k) + \frac{h^2}{2}[2x_k\sin(y_k) + x_k^4\cos(y_k)\sin(y_k)]$$

Можно продолжать считать производные и получать более точный результат.

Замечание 13.2. Заметим, что здесь нет речи о разностной схеме, так как формально достаточно затруднительно посчитать *глобальную* погрешность аппроксимации. Если при первом шаге мы считаем производную в известной нам начальной точке, то при каждом следующем шаге мы будем искать производную в точке, которая никак не привязана к изначальному уравнению.

Методы Адамса первого порядка точности

Решаем задачу Коши $\begin{cases} y'(x) = f(x,y) \\ y(0) = y^0 \end{cases}$

Основная идея берется из точного равенства $y'(x_{k+1}) \equiv y(x_k) + \int_{x_k}^{x_{k+1}} y'(x)dx = y(x_k) + \int_{x_k}^{x_{k+1}} f(x,y)dx$. Различные методы Адамса возникают из разных способов взять интеграл

Лемма 13.1. *Пусть* $g \in C^k[a,b],\ I(g) = \int_a^b g(x)dx$, *тогда*

$$1)\ |I(g) - g(a)(b-a)| \le \|g'\|\frac{(b-a)^2}{2}$$

$$2)\ |I(g) - g(b)(b-a)| \le \|g'\|\frac{(b-a)^2}{2}$$

$$3)\ \left|I(g) - g\left(\frac{a+b}{2}\right)(b-a)\right| \le \|g''\|\frac{(a-b)^3}{24}$$

$$4)\ \left|I(g) - \frac{g(a)+g(b)}{2}(b-a)\right| \le \|g''\|\frac{(b-a)^3}{12}$$

Доказательство. Первые три неравенства берутся из разложения подынтегральной функции в ряд Тейлора с остаточным членом в форме Лагранжа в указанных точках:

$$1) \left| \int_a^b g(a) + g'(\xi)(x-a)dx - g(a)(b-a) \right| = \left| \int_a^b g'(\xi)(x-a)dx \right| \le \|g'\| \int_a^b (x-a)dx \le \|g'\| \frac{(b-a)^2}{2}$$

$$2) \left| \int_a^b g(b) + g'(\xi)(x-b)dx - g(b)(b-a) \right| = \left| \int_a^b g'(\xi)(x-b)dx \right| \le \|g'\| \int_a^b (b-x)dx \le \|g'\| \frac{(b-a)^2}{2}$$

$$3) \left| \int_a^b g\left(\frac{a+b}{2}\right) + g'\left(\frac{a+b}{2}\right)\left(x - \left(\frac{a+b}{2}\right)\right) + \frac{g''(\xi)}{2}\left(x - \left(\frac{a+b}{2}\right)\right)^2 dx - g\left(\frac{a+b}{2}\right)(b-a) \right| =$$

$$\left| \int_a^b g\left(\frac{a+b}{2}\right) dx + \underbrace{\int_a^b g'\left(\frac{a+b}{2}\right)\left(x - \left(\frac{a+b}{2}\right)\right) dx}_{=0,\ \text{(симм. отн. середины)}} + \int_a^b \frac{g''(\xi)}{2}\left(x - \left(\frac{a+b}{2}\right)\right)^2 dx - g\left(\frac{a+b}{2}\right)(b-a) \right|$$

$$= \left| \int_a^b \frac{g''(\xi)}{2}\left(x - \left(\frac{a+b}{2}\right)\right)^2 dx \right| \le \frac{\|g''\|}{2} \left| \int_a^b \left(x - \left(\frac{a+b}{2}\right)\right)^2 dx \right| = \|g''\| \frac{(a-b)^3}{24}$$

Для доказательства 4ого пункта рассматриваем следующий интеграл

$$\frac{1}{2}\int_a^b g''(x)(x-a)(x-b)dx = \frac{1}{2}g'(x)(x-a)(x-b)\Big|_a^b - \frac{1}{2}\int_a^b g'(x)((x-b)+(x-a))dx =$$

$$= -\frac{1}{2}g(x)((x-b)+(x-a))\Big|_a^b + \int_a^b g(x)dx = \int_a^b g(x)dx - \frac{g(a)+g(b)}{2}(b-a)$$

Доказали эквивалентность предложенного интеграла, оценим его:

$$\left| \frac{1}{2}\int_a^b g''(x)(x-a)(x-b)dx \right| \le \frac{\|g''\|}{2} \left| \int_a^b g''(x)(x-a)(x-b)dx \right| \le \frac{\|g''\|}{12}(b-a)^3$$

□

Явный метод Эйлера

Разностная схема получается из пункта 1) доказанной выше леммы:

$$\frac{y_{k+1}-y_k}{h} = f(x_k, y_k) + \underline{\underline{\mathcal{O}}}(h),\ y_0 = y(x_0)$$

Локальная погрешность $O(h^2)$, глобальная погрешность и сходимость $O(h)$.

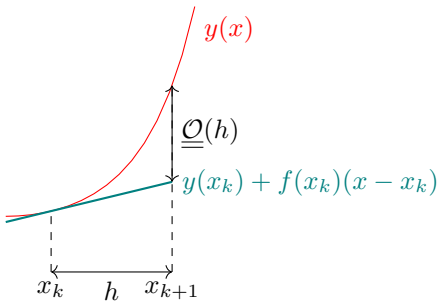


Рис. 7: Пример явного метода Эйлера

Неявный метод Эйлера

Разностная схема получается из пункта 2) доказанной выше леммы:

$$\frac{y_{k+1}-y_k}{h} = f(x_{k+1}, y_{k+1}) + \underline{\underline{\mathcal{O}}}(h),\ y_0 = y(x_0)$$

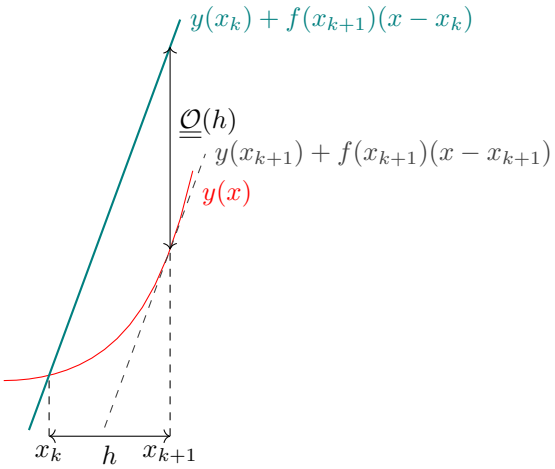


Рис. 8: Пример неявного метода Эйлера

Локальная погрешность $O(h^2)$, глобальная погрешность и сходимость $O(h)$. Является более устойчивой схемой, чем явный метод Эйлера.

Неявный метод называется именно так из-за того, что разностное уравнение в схеме является нелинейным. Для того, чтобы его решать предлагается вводить внутренний итерационный процесс

$$y_{k+1}^{j+1} = y_k + hf(x_{k+1}, y_{k+1}^j),\ y_{k+1}^0 = y_k$$

При достаточно малых h и достаточно гладкой f можно показать, что данное отображение является сжимающим.

Методы Адамса второго порядка точности

Через формулу прямоугольника по средней точке

Из пункта 3) доказанной выше леммы получили следующую оценку

$$y(x_k + h) = y(x_k) + hf(y(x_k + h/2), x_k + h/2) + \underline{\underline{\mathcal{O}}}(h^3)$$

Но в пространстве Y_h не определен узел $x_k + h/2$, поэтому для расчетной формулы нужно думать что-то другое.

Один из вариантов: представить $y(x_k + h/2) := \frac{y(x_k)+y(x_k+h)}{2}$ и воспользоваться снова итерационным процессом для расчетной формулы $y_{k+1} = y_k + hf(\frac{y_k+y_{k+1}}{2}, x_k + h/2) + \underline{\underline{\mathcal{O}}}(h^3)$.

Рассмотрим другой вариант: пусть $\exists\ y_{k+1/2}^* \overset{\text{def}}{=} y_k + \underline{\underline{\mathcal{O}}}(h^2)$, тогда мы можем записать расчетную формулу следующим образом:

$$\begin{aligned} y_{k+1} = y_k + hf(y_k, x_k + h/2) \pm hf(x_k + h/2, y_{k+1/2}^*) + \underline{\underline{\mathcal{O}}}(h^3) = \\ = y_k + hf(x_k + h/2, y_{k+1/2}^*) + (hf(y_k, x_k + h/2) - hf(x_k + h/2, y_{k+1/2}^*)) + \underline{\underline{\mathcal{O}}}(h^3) \end{aligned}$$

Воспользуемся теоремой о среднем: $\tilde{y_k}$ - точка между $y_{k+1/2}^*$ и y_k

$$y_{k+1} = y_k + hf(x_k + h/2, y_{k+1/2}^*) + \frac{h}{2}(f_x(\tilde{y_k}, x_k + h/2)) \underbrace{(y_{k+1/2}^* - y_k)}_{\underline{\underline{\mathcal{O}}}(h^2)} + \underline{\underline{\mathcal{O}}}(h^3) = y_k + hf(x_k + h/2, y_{k+1/2}^*) + \underline{\underline{\mathcal{O}}}(h^3)$$

Точку $y_{k+1/2}^*$ можно посчитать с помощью явного метода Эйлера, который как раз и обеспечивает второй порядок сходимости $y_{k+1/2}^* = y_k + \frac{h}{2}f(x_k, y_k)$ Итоговая расчетная форумла

$$\begin{cases} y_{k+1/2}^* = y_k + \frac{h}{2}f(x_k, y_k) \\ y_{k+1} = y_k + hf(x_k + h/2, y_{k+1/2}^*) \\ y_0 = y(x_0) \end{cases}$$

Через точку x_k проводится касательная с углом наклона, равным коэффициенту касательной, построенной по явному методу Эйлера через среднюю точку.

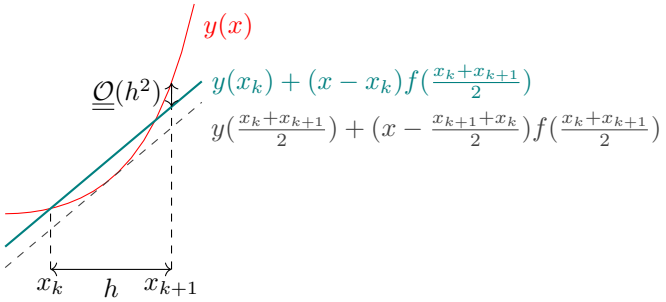


Рис. 9: Пример неявного метода Адамса через среднюю точку

Через формулу трапеции

Разностная схема получается из пункта 4) доказанной выше леммы:

$$y_{k+1} = y_k + h\frac{f(x_k, y_k) + f(x_{k+1}, y_{k+1})}{2} + \underline{\underline{\mathcal{O}}}(h^3)$$

Схема получается нелинейная, для решения можно воспользоваться итерационным методом. Мы воспользуемся аналогично формуле прямоугольника по центральной точке хитростью.

Пусть $\exists\ y^*_{k+1} \stackrel{\text{def}}{=} y_k + \underline{\underline{\mathcal{O}}}(h^2)$, тогда мы можем записать расчетную формулу следующим образом:

$$y_{k+1} = y_k + h\frac{f(x_k, y_k) + f(x_{k+1}, y_{k+1})}{2} \pm \frac{h}{2}f(x_{k+1}, y^*_{k+1}) + \underline{\underline{\mathcal{O}}}(h^3) =$$
$$= y_k + h\frac{f(x_k, y_k) + f(x_{k+1}, y^*_{k+1})}{2} + \frac{h}{2}(f(x_{k+1}, y_{k+1}) - f(x_{k+1}, y^*_{k+1})) + \underline{\underline{\mathcal{O}}}(h^3)$$

Воспользуемся теоремой о среднем: $\tilde{y_k}$ - точка между y^*_{k+1} и y_k

$$y_{k+1} = y_k + h\frac{f(x_k, y_k) + f(x_{k+1}, y^*_{k+1})}{2} + \frac{h}{2}f_x(x_{k+1}, \tilde{y_k})\underbrace{(y^*_{k+1} - y_k)}_{\underline{\underline{\mathcal{O}}}(h^2)} + \underline{\underline{\mathcal{O}}}(h^3) =$$
$$= y_k + h\frac{f(x_k, y_k) + f(x_{k+1}, y^*_{k+1})}{2} + \underline{\underline{\mathcal{O}}}(h^3)$$

Точку y^*_{k+1} будем искать с помощью явного метода Эйлера $y^*_{k+1} = y_k + hf(x_k, y_k)$. Итоговая расчетная формула

$$\begin{cases} y^*_{k+1} = y_k + hf(x_k, y_k) \\ y_{k+1} = y_k + h\frac{f(x_k,y_k)+f(x_{k+1},y^*_{k+1})}{2} + \underline{\underline{\mathcal{O}}}(h^3) \\ y_0 = y(x_0) \end{cases}$$

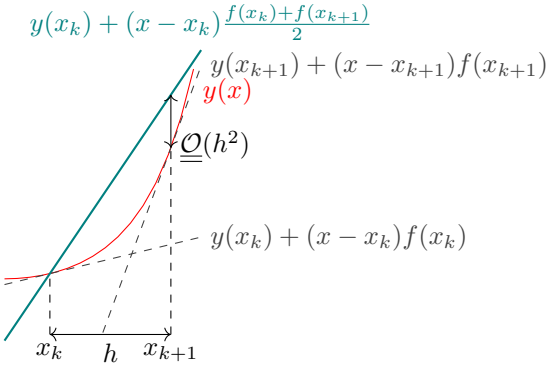


Рис. 10: Пример неявного метода Адамса через формулу трапеции

Через точку x_k проводится касательная с углом наклона, равным среднему арифметическому коэффициентов касательных, построенных по явному и неявному методам Эйлера.

Замечание 13.3. Предложенные варианты с заменой на явные методы Эйлера облегчают подсчет итоговой расчетной формулы, но при этом отрицательно влияют на устойчивость задачи.

14 Методы Рунге–Кутта для решения задачи Коши.

Решая задачу $y'(x) = f(x, y(x)),\ y(x_0) = y_0$, хотим иметь возможность точно перейти от $y(x) \to y(x+h)$ с наибольшим порядком точности $\underline{\underline{\mathcal{O}}}(h^{s+1})$.

Основная идея заключается в том, что в отличие от явного метода Эйлера, который использует значение производной только в одной точке, мы возьмем несколько значений и с помощью весов добьемся более высокого порядка аппроксимации.

Зафиксируем некоторые числа

$$\alpha_2, \dots, \alpha_q,\ p_1, \dots, p_q,\ \beta_{ij},\ 0 < j < i \le q$$

Последовательно вычислим

$$\begin{aligned}
k_1(h) &= hf(x, y(x)) \\
k_2(h) &= hf(x + \alpha_2 h, y(x) + \beta_{21} k_1(h)) \\
k_3(h) &= hf(x + \alpha_3 h, y(x) + \beta_{32} k_2(h) + \beta_{31} k_1(h)) \\
&\vdots \\
k_q(h) &= hf(x + \alpha_q h, y(x) + \sum_{i=1}^{q-1} \beta_{qi} k_i(h))
\end{aligned}$$

Тогда расчетная формула будем иметь вид:

$$y(x+h) \simeq z(x+h) = y(x) + \sum_{i=1}^{q} p_i k_i(h)$$

Обозначим погрешность метода на шаге как $\varphi(h) = y(x+h) - z(x+h)$. Если у $f(x, y)$ существует s производных, то верна формула Тейлора, запишем ее с остаточным членом в форме Лагранжа

$$\varphi(h) = \sum_{i=0}^{s} \frac{\varphi^{(i)}(0)}{i!} h^i + \frac{\varphi^{(s+1)}(\theta h)}{(s+1)!} h^{s+1},\ 0 \le \theta \le 1$$

Осталось подобрать α_i, p_i, β_{ij} так, чтобы $\varphi^{(i)}(0) = 0,\ i = 0, \dots, s$. Если нам удастся это сделать, то локальная погрешность метода составляет $\underline{\underline{\mathcal{O}}}(h^{s+1})$, и величина s называется *порядком* метода.

Пример 14.1. Выпишем формулы расчета методом Рунге-Кутта с $q = 1$.

$$\varphi(h) = y(x+h) - y(x) - p_1 k_1(h) = y(x+h) - y(x) - p_1 hf(x, y)$$

Хотим занулить как можно больше слагаемых, то есть производных в 0. За счет выбора φ сразу известно, что $\varphi(0) = 0$. Посмотрим на следующие слагаемые подробнее:

$$\varphi'(0) = (y(x+h) - y(x) - p_1 hf(x, y))'|_{h=0} = (y'(x+h) - p_1 f(x, y))|_{h=0} = y'(x) - p_1 f(x, y) = (1 - p_1) f(x, y)$$

$$\varphi''(0) = (y(x+h) - y(x) - p_1 hf(x, y))''|_{h=0} = y''(x)$$

Чтобы достигнуть локальной погрешности $\underline{\underline{\mathcal{O}}}(h^2)$ возможно взять только $p_1 = 1$. Погрешность в таком случае равна $\varphi(h) = \frac{h^2}{2} y''(x + \theta h)$.

Итоговая расчетная формула имеет вид

$$y(x+h) \simeq y(x) + hf(x, y)$$

То есть формула Рунге-Кутта с $q = 1$ совпадает с явной формулой Эйлера.

Пример 14.2. Посчитаем все формулы Рунге-Кутта для $q = 2$.

$$\varphi(h) = y(x+h) - y(x) - p_1 k_1(h) - p_2 k_2(h) = y(x+h) - y(x) - p_1 k_1(h) - p_2 hf(\hat{x}, \hat{y})$$

где $\hat{x} = x + \alpha_2 h,\ \hat{y} = y(x) + \beta_{21} k_1(h)$.

$$\varphi(0) = 0;\ \varphi'(0) = (y(x+h) - y(x) - p_1 k_1(h) - p_2 k_2(h))'_{h=0} =$$

$$\left| \begin{array}{c} (p_1 k_1(h))' = p_1 f(x, y) \\ (p_2 k_2(h))' = p_2 f(\hat{x}, \hat{y}) + p_2 h \underbrace{(\alpha_2 f_x(\hat{x}, \hat{y}) + \beta_{21} f(x, y) f_y(\hat{x}, \hat{y}))}_{q(h)} \end{array} \right|$$

$$= (y'(x+h) - p_1 f(x, y) - p_2 f(\hat{x}, \hat{y}) + p_2 h q(h))_{h=0} =$$

$$= y'(x) - p_1 f(x, y) - p_2 f(x, y) = f(x, y)(1 - p_1 - p_2)$$

$$\Rightarrow \varphi'(0) = 0 \Leftrightarrow 1 - p_1 - p_2 = 0 \tag{1}$$

$$\varphi''(0) = (y(x+h) - y(x) - p_1 k_1(h) - p_2 k_2(h))''_{h=0} =$$
$$\left| \begin{array}{c} (p_1 k_1(h))'' = 0 \\ (p_2 k_2(h))'' = (p_2 f(\hat{x},\hat{y}) + p_2 h q(h))' = 2p_2 q(h) + p_2 h q'(h) = \\ = 2p_2 q(h) + p_2 h(\alpha_2^2 f_{xx}(\hat{x},\hat{y}) + 2\alpha_2\beta_{21} f(x,y) f_{xy}(\hat{x},\hat{y}) + \beta_2^2 f^2(x,y) f_{yy}(\hat{x},\hat{y})) \end{array} \right|$$
$$= \big(y''(x+h) - 2p_2 q(h) - p_2 h(\alpha_2^2 f_{xx}(\hat{x},\hat{y}) + 2\alpha_2\beta_{21} f_{xy}(\hat{x},\hat{y}) + \beta_2^2 f^2(x,y) f_{yy}(\hat{x},\hat{y}))\big)_{h=0} =$$
$$\left| \; y''_{hh} = (y'_h)'_h = (f)' = f_x + f_y y'_h = f_x + f_y f \; \right|$$
$$= (1 - 2p_2\alpha_2) f_x(x,y) + (1 - 2p_2\beta_{21}) f_y(x,y) f(x,y)$$

$$\Rightarrow \varphi''(0) = 0 \Leftrightarrow \begin{cases} 1 - 2p_2\alpha_2 = 0 \\ 1 - 2p_2\beta_{21} = 0 \end{cases} \tag{2}$$

$$\varphi'''(0) = (y(x+h) - y(x) - p_1 k_1(h) - p_2 k_2(h))'''_{h=0} =$$
$$\left| \begin{array}{c} (p_1 k_1(h))''' = 0 \\ (p_2 k_2(h))''' = (2p_2 q(h) + p_2 h q'(h))' = 3p_2 q'(h) + \underline{\underline{\mathcal{O}}}(h) \end{array} \right|$$
$$= \big(y'''(x+h) - 3p_2(\alpha_2^2 f_{xx}(\hat{x},\hat{y}) + 2\alpha_2\beta_{21} f_{xy}(\hat{x},\hat{y}) + \beta_2^2 f^2(x,y) f_{yy}(\hat{x},\hat{y})) + \underline{\underline{\mathcal{O}}}(h)\big)_{h=0} =$$
$$\left| \; y'''_{hhh} = (y''_h)'_h = (f_x + f_y f)' = f_{xx} + f_{xy} f + (f_{yx} + f_{yy} f) f + f_y (f_x + f_y f) = f_{xx} + 2f_{xy} f + f_{yy} f^2 + f_y y'' \; \right|$$
$$= (1 - 3p_2\alpha_2^2) f_{xx} + (2 - 6\alpha_2\beta_{21}) f_{xy} f + (1 - 3p_2\beta_2^2) f_{yy} f^2 + f_y y''$$

Если в исходной задаче будет дано y такое, что $y'' \neq 0$, то это соотношение никогда не обратится в 0. Отсюда следует, что в общем случае нельзя построить формулу Рунге-Кутта со значениями $s = 3$, $q = 2$.

Рассмотрим конкретные примеры, удовлетворяющие соотношениям (1) и (2):

1. $p_1 = 0,\ p_2 = 1,\ \alpha_1 = \beta_{21} = 1/2$. Итоговая расчетная формула имеет вид

$$y(x+h) \simeq y(x) + h f(x + \frac{h}{2}, y(x) + \frac{h}{2} f(x,y)))$$

Получили неявный метод Адамса второго порядка через формулу прямоугольника по средней точке.

2. $p_1 = p_2 = 1/2,\ \alpha_2 = \beta_{21} = 1$. Итоговая расчетная формула имеет вид

$$y(x+h) \simeq y(x) + \frac{h}{2}(f(x,y) + f(x+h, y(x) + h f(x,y)))$$

Получили неявный метод Адамса второго порядка через формулу трапеции.

15 Вычисление главного члена погрешности для простейших схем для задачи Коши. Оценка глобальной погрешности явного одношагового метода.

Вычисление главного члена погрешности

Получили разностную схему, хотим узнать насколько она пригодна. Предлагается рассмотреть модельную дифференциальную задачу и модельное разностное решение.

Рассматривается задача Коши $\begin{cases} y' = y,\ x \in [0,1] \\ y(0) = 1 \end{cases}$ с точным решением e^x. В качестве примера рассмотрим явную схему Эйлера $\begin{cases} \frac{y_{k+1}-y_k}{h} = y_k,\ h = \frac{1}{N} \\ y_0 = 1 \end{cases}$ с решением $y_{k+1} = (1+h)y_k = (1+h)^k$.

Хотим узнать отличие в последней точке $x_N = 1$: $y(x_N) - y_N = c_0 + c_1 \cdot h + c_2 \cdot h^2 + \dots$

$$y_N = (1+h)^N = (1+h)^{\frac{1}{h}} = \exp\left\{\ln(1+h)^{\frac{1}{h}}\right\} = \exp\left\{\frac{1}{h}\ln(1+h)\right\} \underset{h\to 0}{=}$$
$$= \exp\left\{\frac{1}{h}\left(h - \frac{h^2}{2} + \underline{\underline{\mathcal{O}}}(h^3)\right)\right\} = \exp\left\{1 - \frac{h}{2} + \underline{\underline{\mathcal{O}}}(h^2)\right\} = e \cdot \exp\left\{-\frac{h}{2} + \underline{\underline{\mathcal{O}}}(h^2)\right\} = e \cdot \left(1 - \frac{h}{2} + \underline{\underline{\mathcal{O}}}(h^2)\right)$$

$$\Rightarrow y(x_N) - y_N \underset{h\to 0}{=} e - e \cdot \left(1 - \frac{h}{2} + \underline{\underline{\mathcal{O}}}(h^2)\right) = e \cdot \frac{h}{2} + \underline{\underline{\mathcal{O}}}(h^2)$$

Таким образом $c_0 = 0,\ c_1 = \frac{e}{2}$ - главный член погрешности.

Оценка глобальной погрешности явного одношагового метода

Рассматривается задача Коши $y'(x) = f(x,y),\ y(x_0) = y^0$. Для ее решения выбирается явный одношаговый метод $y_{k+1} = F(y_k, x_k, x_{k+1} - x_k),\ k = 0, \dots, n-1,\ x_n = x_0 + X$ с локальной погрешностью $s+1$ порядка. Нас интересует итоговая погрешность в n точке, то есть величина $|y_n - y(x_n)|$. Докажем вспомогательную лемму.

Лемма 15.1 (Гронуолл)**.** *Пусть $y_{(1)}(x)$, $y_{(2)}(x)$ - два решения задачи $y'(x) = f(x,y),\ x \in [a,b]$, пусть f непрерывна и непрерывно дифференцируема по y. Тогда верное следующее выражение*

$$y_{(1)}(b) - y_{(2)}(b) = (y_{(1)}(a) - y_{(2)}(a)) \exp\left(\int_a^b f_y(\tau, \tilde{y}(\tau))d\tau\right)$$

где $\tilde{y}$ заключена между $y_{(1)}$ и $y_{(2)}$.

Доказательство. Из линейности дифференциального оператора следует следующее

$$(y_{(1)}(x) - y_{(2)}(x))' = f(x, y_{(1)}) - f(x, y_{(2)})$$

Так как f непрерывна и непрерывно дифференцируема по y, то по теореме о среднем для f:

$$(y_{(1)}(x) - y_{(2)}(x))' = f_y(x, \tilde{y})(y_{(1)}(x) - y_{(2)}(x)) \tag{1}$$

Так как функция $\tilde{y}$ фиксирована, то есть $f_y(x, \tilde{y})$ зависит только от x, то обозначим $g(x) \overset{\text{def}}{=} f_y(x, \tilde{y})$.

Делаем трюк: домножаем (1) и слева и справа на $\exp\left(\int_a^x g(\tau)d\tau\right)$:

$$\exp\left(\int_a^x g(\tau)d\tau\right)(y_{(1)}(x) - y_{(2)}(x))' = g(x)(y_{(1)}(x) - y_{(2)}(x)) \exp\left(\int_a^x g(\tau)d\tau\right) =: (\star)$$

Посчитаем следующее выражение

$$\frac{d}{dx}\left[\exp\left(-\int_a^x g(\tau)d\tau\right)(y_{(1)}(x) - y_{(2)}(x))\right] =$$
$$\underbrace{\underbrace{\frac{d}{dx}\left[-\int_a^x g(\tau)d\tau\right]}_{-g(x)} \exp\left(-\int_a^x g(\tau)d\tau\right)(y_{(1)}(x) - y_{(2)}(x))}_{-(\star)} + \underbrace{\exp\left(-\int_a^x g(\tau)d\tau\right)(y_{(1)}(x) - y_{(2)}(x))'}_{(\star)} = 0$$

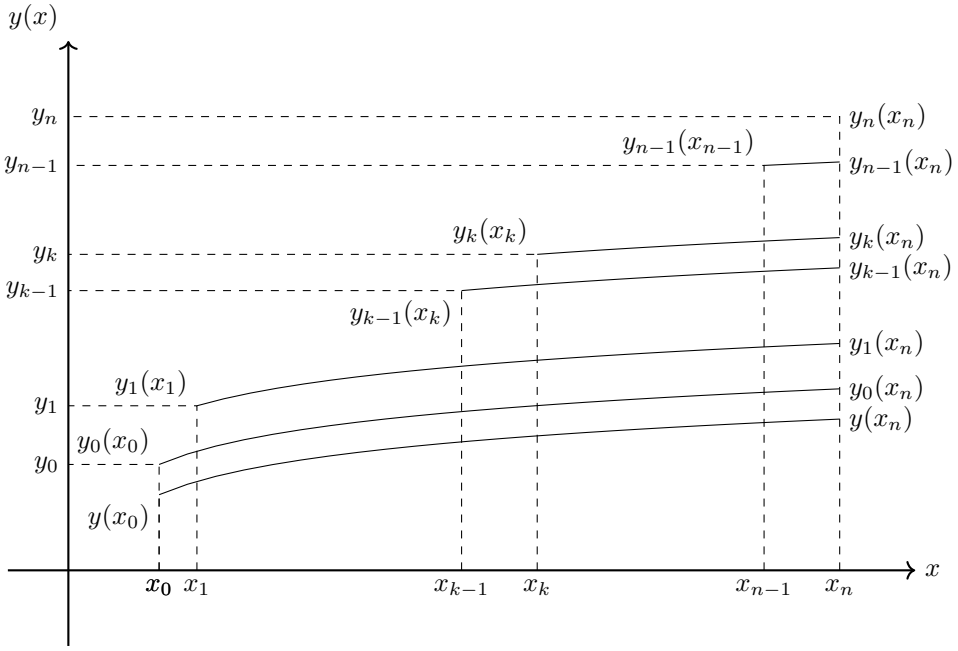
Проинтегрируем полученное тождество по отрезку $[a, b]$:

$$\int_a^b \frac{d}{dx}\left[\exp\left(-\int_a^x g(\tau)d\tau\right)(y_{(1)}(x) - y_{(2)}(x))\right]dx = \int_a^b 0\ dx$$

$$\exp\left(-\int_a^b g(\tau)d\tau\right)(y_{(1)}(b) - y_{(2)}(b)) - \exp\left(-\int_a^a g(\tau)d\tau\right)(y_{(1)}(a) - y_{(2)}(a)) = 0$$

$$y_{(1)}(b) - y_{(2)}(b) = (y_{(1)}(a) - y_{(2)}(a))\exp\left(\int_a^b f_y(\tau, \tilde{y}(\tau))d\tau\right)$$

□



Рассмотрим рисунок.

- Начальное условие $y(x_0)$ порождает интегральную кривую y, которая даст нам точное решение в точке x_n.
- После внесения $y(x_0)$ в компьютер мы получаем значение $y_0 =: y_0(x_0)$. Мы вносим погрешность в виде машинного нуля: $r_0 = y(x_0) - y_0$. Начальное условие y_0 породит интегральную кривую $y_0(x)$ со значением на конце отрезка $y_0(x_n)$.
- Явная формула подсчета следующего значения $y_{k+1} = F(x_k, y_k, x_{k+1} - x_k)$ позволяет узнать значение в точке x_1 с какой-то точностью: $y_1 - y_0(x_1) = c_1 h_1^{s+1} + \delta_1$, где δ_1 - погрешность вычислений первого шага, $h_1 := x_1 - x_0$.
- Используя формулу подсчета мы можем подсчитать значения в каждой из точек x_k. Обратим внимание, что каждая интегральная кривая является решением исходной задачи со своими начальными данными $\begin{cases} y' = f(x, y) \\ y(x_k) = y_k \end{cases}$

Таким образом мы можем явно выписать чему равна искомая величина

$$y_n - y(x_n) = \sum_{k=1}^{n}(y_k(x_n) - y_{k-1}(x_n)) + y_0(x_n) - y(x_n)$$

Так как каждая из наших интегральных кривых является решением на отрезках $[x_k, x_n]$, мы можем воспользоваться леммой Гронуолла:

$$y_n - y(x_n) = \sum_{k=1}^{n}(y_k(x_k) - y_{k-1}(x_k))\exp\left[\int_{x_k}^{x_n} f_y(\tau, \tilde{y}(\tau))d\tau\right] + r_0\exp\left[\int_0^{x_n} f_y(\tau, \tilde{y}(\tau))d\tau\right]$$

Так как локальная оценка погрешности нам известна, то мы можем переписать полученное выражение:

$$y_n - y(x_n) = \sum_{k=1}^{n} (c_k h_k^{s+1} + \delta_k) \exp\left[\int_{x_k}^{x_n} f_y(\tau, \tilde{y}(\tau)) d\tau\right] + r_0 \exp\left[\int_0^{x_n} f_y(\tau, \tilde{y}(\tau)) d\tau\right] \tag{2}$$

Получили довольно сложную формулу, покажем, какие выводы с помощью можно сделать.
Пусть $h_k \le h$, $\delta_k \le \delta$, $c_k \le c$.

1. Пусть $0 < |f_y| \le L < \infty$. Тогда выражение (2) можно оценить:

$$|y_n - y(x_n)| \le \sum_{k=1}^{n} (ch^s h + \delta) \exp\left[L(x_n - x_k)\right] + r_0 \exp\left[XL\right] \le \exp\left[LX\right] (ch^s X + n\delta + r_0)$$

   Таким образом, чем больше длина отрезка, тем больше итоговая погрешность, и растет она экспоненциально! Из второго слагаемого напрашивается вывод: чем больше шагов мы делаем, тем больше становится итоговая вычислительная погрешность.

2. Более занимательный результат получается, если f является убывающей, то есть $f_y \le -L \le 0$

$$|y_n - y(x_n)| \le (ch^s h + \delta) \sum_{k=1}^{n} \exp\left[-Lkh\right] + r_0 \exp\left[-XL\right] \le \frac{ch^s h + \delta}{1 - \exp[-Lh]} + r_0 \exp\left[-XL\right] \underset{h<<L,\ \text{Тейлор}}{\le}$$
$$\frac{ch^s}{L} + \frac{\delta n}{XL} + r_0 \exp\left[-XL\right]$$

   Таким образом получаем, что чем сильнее убывает правая часть, тем точнее получаем решение. От количества шагов вычислительная погрешность растет, но не с экспоненциальной скоростью, как в предыдущем примере.

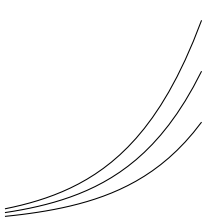
16 Устойчивые и неустойчивые задачи. Жесткие системы.

Устойчивые и неустойчивые задачи

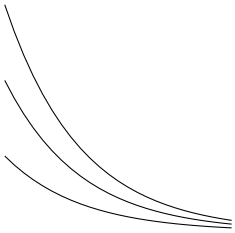
Рассматривается поведение решений следующих задач

$$\begin{array}{lll} y_1' = y & y_2' = -y_2 & y_3' = -2(x-1)y_3 \\ y_1(x) = ce^x & y_2(x) = ce^{-x} & y_3' = ce^{-(x-1)^2} \end{array}$$

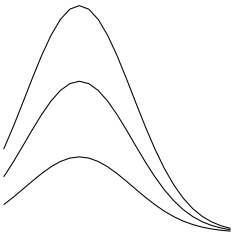
Интегральные кривые первого семейства расходятся с увеличением x, второго сближаются, а третьего сначала расходятся, а затем сближаются. Так как на k-ом шаге найденное приближенное значение y_k смещается с интегральной кривой $y_{k-1}(x)$, то внесенная в результате этого погрешность может в зависимости от поведения семейства решений либо возрастать, либо уменьшаться, см. рис.:



Ce^x



Ce^{-x}



$Ce^{-(x-1)^2}$

В связи с этим имеется существенная разница в численном интегрировании двух на первый взгляд эквивалентных задач

$$\begin{cases} y_1' = y_1 \\ y_1(0) = 1 \end{cases} \qquad \begin{cases} y_2' = e^x \\ y_2(0) = 1 \end{cases}$$

так как соответствующие им семейства интегральных кривых существенно отличаются: $y(x) = ce^x$, $y(x) = c + e^x$.

Выясним, какое из рассмотренных уравнение явный метод Эйлера проинтегрирует точнее. Пусть при внесении начальных данных внесется погрешность $\delta_0 > 0$, то есть $y_0 = y(0) + \delta_0$

1. Посчитаем глобальную погрешность для первой задачи:

$$\frac{y_{k+1} - y_k}{h} = y_k \Rightarrow y_n = (1+h)y_{n-1} = (1+h)^{1/h} y_0 = (1+h)^{1/h}(y(0) + \delta_0) = (1+h)^{1/h}(1+\delta_0)$$

$$(1+h)^{\frac{1}{h}} = \exp\left\{\frac{1}{h}\ln(1+h)\right\} = \exp\left\{1 - \frac{h}{2} + \underline{\underline{\mathcal{O}}}(h^2)\right\} = e \cdot \left(1 - \frac{h}{2} + \underline{\underline{\mathcal{O}}}(h^2)\right)$$

$$|y(x_n) - y_n| = \left| e - e \cdot \left(1 - \frac{h}{2} + \underline{\underline{\mathcal{O}}}(h^2)\right)(1+\delta_0)\right| = e\frac{h}{2} + e\delta_0\left(1 + \frac{h}{2}\right) + \underline{\underline{\mathcal{O}}}(h^2)$$

2. Посчитаем глобальную погрешность для второй задачи:

$$\frac{y_{k+1} - y_k}{h} = e^{x_k} \Rightarrow y_n = y_{n-1} + he^{x_{n-1}} = \sum_{i=0}^{n-1} he^{x_i} + \delta_0$$

Полученная расчетная формула соответстует составной формуле прямоугольника по крайней точке. Для такой формулы позднее будет получена оценка

$$\left| y(x_n) - \sum_{i=0}^{n-1} he^{x_i} - \delta_0 \right| \le \left| y(x_n) - \sum_{i=0}^{n-1} he^{x_i} \right| + \delta_0 \le \|(e^x)'\|_{x\in[0,1]} \frac{(1-0)^2}{2N} + \delta_0 = e\frac{h}{2} + \delta_0$$

Вывод: вторая схема проинтегрирует точнее. Стоит обратить внимание, что погрешность вносится не только при задании y_0, но и при каждом вычислении. В случае первой задачи такая погрешность будет расти экспоненциально, в отличие от второй задачи.

Жесткие схемы

Рассмотрим задачу интегрирования уравнения $y' = -ay,\ a \equiv \text{const} > 0,\ x \in [0, X]$. Для решения данной задачи запишем явную разностную схему

$$\frac{y_{k+1} - y_k}{h} = -ay_k \Rightarrow y_{k+1} = (1 - ha)y_k$$

Решение исходной дифференциальной задачи убывает при увеличении x. Логично требовать, чтобы решение разностной задачи тоже обладало этим же качеством: $|1 - ha| < 1$. Соответствующее множество шагов h называется *областью устойчивости* разностной схемы, а максимально допустимый шаг $h_{cou} = 2/a$ - *числом Куранта*.

Так как $y(x + h) = y(x) + hy'(x) + \frac{h^2}{2}y''(\xi),\ x \le \xi \le x + h$, то погрешность аппроксимации имеет вид $\left|\frac{h^2}{2}y''(\xi)\right|$. Величина $y''(\xi) = a^2e^{-a\xi} \ll 1$ при $a\xi \gg 1$. Поэтому при $a\xi \gg 1$ для достижения требуемой точности локальной аппроксимации $|hy''(\xi)| \le \varepsilon$ не требуются мелкие шаги, но $h \le h_{cou}$ для всех x из условия качественного совпадения решений (условия устойчивости).

Данный пример показывает, что разумно выделить класс так называемых *жестких задач*. Будем считать задачу $y' = f$ жесткой, если *характерное время изменения решения много меньше отрезка интегрирования* (в данном случае это означает, что $aX \gg 1$). Система уравнений $y = Ay, y = (y_1, \ldots, y_n)^T$ считается жесткой, если

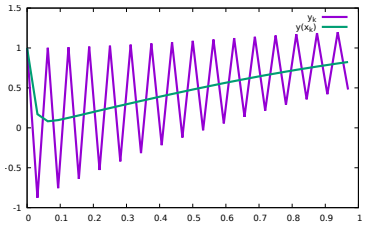
$$1)\ \text{Re}(\lambda_i(A)) > 0 \quad 2)\ s = \frac{\max_i |\text{Re}(\lambda_i(A))|}{\min_i |\text{Re}(\lambda_i(A))|} \gg 1$$

Число s принято называть *числом жесткости*. Данное определение обобщается на случай матриц $A(x)$.

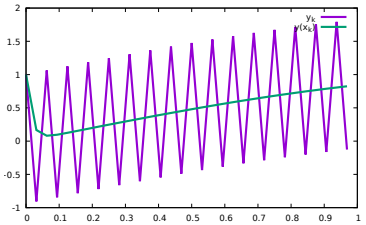
В случае одного уравнения задача будет жесткой, если решение содержит несколько компонент с существенно отличающимися характерными временами изменения. Рассмотрим задачу

$$y' = -a(y - \sin x) + \cos x,\ y(0) = 1,\ a \gg 1 \Rightarrow y(x) = e^{-ax} + \sin x$$

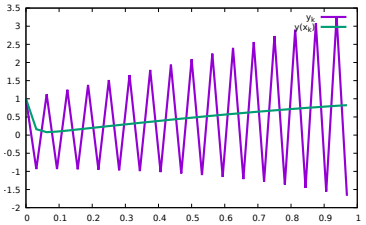
Здесь можно выделить пограничный слой быстрого изменения решения, далее решение мало отличается от плавной функции $\sin x$. Однако $\forall\ x$ необходимо выполнение условия $h \le h_{cou}$.



$N = 32,\ X = 1,\ A = 63$



$N = 32,\ X = 1,\ A = 64$



$N = 32,\ X = 1,\ A = 65$

Рис. 11: Результаты явной схемы Эйлера с различными входными параметрами

Еще раз отметим, что устойчивость дифференциальной задачи определяется типом уравнения и значениями параметров, устойчивость разностной задачи - типом разностной схемы, значениями параметров и величиной шага. Так явные схемы с постоянным шагом обычно требуют мелких шагов, но просты в реализации. Неявные схемы обычно имеют менее жесткие условия на h за счет дополнительных вычислений в общем случае.

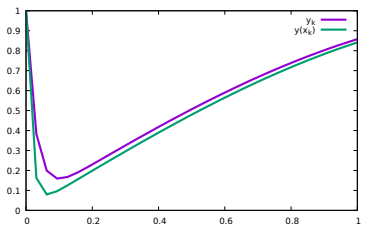


Рис. 12: Результаты неявной схемы Эйлера с $N = 32,\ X = 1,\ A = 65$

17 Метод Лебедева решения жестких систем.

Напомним определение жесткой задачи

Опр. 17.1. Система $y' = Ay$, $y = (y_1, \dots, y_n)^T$ называется жесткой если

$$1)\ \mathrm{Re}(\lambda_i(A)) < 0 \quad 2)\ s = \frac{\max_i |\mathrm{Re}(\lambda_i(A))|}{\min_i |\mathrm{Re}(\lambda_i(A))|} \gg 1$$

Число s называется *числом жесткости*

Было показано, что для таких систем не требуется выбор маленьких шагов h для достижения требуемой точности локальной аппроксимации. При этом взять слишком большие шаги мы не можем, так как должна соблюдаться устойчивость задачи для сохранения сходимости. Хотим научиться подбирать большие шаги так, чтобы сохранялась и устойчиовсть и желаемая аппроксимация.

Рассматривается задача

$$y' = Ay,\ y = (y_1, \dots, y_N)^T,\ A \in \mathbb{R}^{N\times N}$$

Для нее выбирается явная схема с переменными шагами. Тогда за N шагов получим

$$y_N = y_{N-1}(1 - h_N A) = y_0 \underbrace{\prod_{k=1}^{N}(1 - h_k A)}_{S} = y_0 P_N(A)$$

Лемма 17.1. *Пусть собственные векторы матрицы A образуют базис. Тогда*

$$\overline{\lim_{k\to\infty}} \left\|S^k\right\| \le \mathrm{const} \Leftrightarrow |\lambda_i(S)| \le 1$$

Доказательство. Найдем собственные числа матрицы S. так как собственные вектора матрицы A образуют базис, то вектор $y_0 \in \mathbb{R}^N$ можно разложить по собственным векторам $y_0 = \sum_{k=1}^{N} c_i y^{(k)}$. Подействуем матрицей S на такой вектор

$$Sy_0 = \prod_{k=1}^{N}(I - h_k A)y_0 = \prod_{k=1}^{N}(I - h_k A)\sum_{i=1}^{N} c_i y^{(i)} = \sum_{i=1}^{N}\prod_{k=1}^{N}(I - h_k A)c_i y^{(i)}$$

Посмотрим на последнее выражение подробнее. В каждом из слагаемых участвует полином относительно матрицы A: $\prod_{k=1}^{N}(I - h_k A)y^{(i)} = (1 + \ldots + h_k A^N)y^{(i)} = (1 + \ldots + h_k \lambda_i^N)y^{(i)}$. Значит

$$Sy_0 = \sum_{i=1}^{N}\prod_{k=1}^{N}(1 - h_k\lambda_i)c_i y^{(i)} \Rightarrow \lambda_i(S) = \prod_{k=1}^{N}(1 - h_k\lambda_i(A))$$

Тогда

$$\overline{\lim_{k\to\infty}} \left\|S^k\right\| \le \mathrm{const} \Leftrightarrow \left|\prod_{k=1}^{N}(1 - h_k\lambda_i(A))\right| \le 1$$

□

Из этой леммы следует, что для обеспечения сходимости мы можем брать h больший, чем h_{cou} до тех пор, пока $\left|\prod_{k=1}^{N}(1 - h_k\lambda_i(A))\right| \le 1$.

Мы хотим получить наибольший отрезок интегрирования и при этом сохранить сходимость. Потребуем $\lambda_i \in [0, M]$, и тогда нам удастся переформулировать эти желания в терминах полинома $P_N(x)$:

$$\max_{x\in[0,M]} |P_N(x)| \le 1,\ \sum_{i=1}^{N} h_i = -P_N'(0) \to \sup \tag{1}$$

Мы хотим подобрать коэффициенты полинома - искомые h_k - так, чтобы выполнялись эти два условия. Нам известно, что на отрезке норма многочлена Чебышева ограничена 1, и по теореме Маркова производная многочлена Чебышева вне соответствующего интервала максимальная среди всех остальных полиномов. Многочлен Чебышева определен на $t \in [-1, 1]$, сделаем замену переменных на $x \in [0, M]$:

$$x = \frac{M + 0}{2} + \frac{M - 0}{2}t \Leftrightarrow t = \frac{2x - M}{M}$$

Таким образом, решением задачи (1) является полином $T_N(\frac{2x-M}{M})$. Так как нам известны корни многочлена Чебышева на $[0, M]$, то можем получить формулу для шагов

$$h_k = \left(\frac{M}{2} + \frac{M}{2}\cos\left(\frac{(2k-1)\pi}{2N}\right)\right)^{-1},\ k = 1, \ldots, N$$

Вычислим соответствующую длину отрезка интегрирования, то есть значение $P'_N(0)$

$$T_N(x) = \cos(N \arccos x) \Rightarrow T'_N(x) = \frac{N(\sin(N \arccos x))}{\sqrt{1-x^2}} = \frac{N \sin(N \arccos x)}{\sin \arccos x} = NU_{N-1}(x)$$

Отсюда находим

$$\left|\frac{d}{dx}T_N\left(\frac{2x-M}{M}\right)\right|_{x=0} = \left|\frac{2N}{M}U_{N-1}(x)\right|_{x=-1}$$

Так как подставить -1 сразу не выходит, то посчитаем ассимптотику функции, записав в форме многочлена

$$U_{N-1}(x) = \frac{(x+\sqrt{x^2-1})^N - (x-\sqrt{x^2-1})^N}{2\sqrt{x^2-1}} = x^N \frac{(1+\frac{\sqrt{x^2-1}}{x})^N - (1-\frac{\sqrt{x^2-1}}{x})^N}{2\sqrt{x^2-1}} =$$

$$= x^N \frac{1 + N\frac{\sqrt{x^2-1}}{x} + \underline{\underline{\mathcal{O}}}\left(\left(\frac{\sqrt{x^2-1}}{x}\right)^2\right) - 1 + N\frac{\sqrt{x^2-1}}{x} + \underline{\underline{\mathcal{O}}}\left(\left(\frac{\sqrt{x^2-1}}{x}\right)^2\right)}{2\sqrt{x^2-1}} = Nx^{N-1} + \underline{\underline{\mathcal{O}}}\left(x^{N-2}\sqrt{x^2-1}\right)$$

Таким образом, итоговый отрезок интегрирования

$$\left|\frac{d}{dx}T_N\left(\frac{2x-M}{M}\right)\right|_{x=0} \approx \frac{2N^2}{M}$$

Для сравнения, если пользоваться постоянным шагом, который сохраняет сходимость $\left(h_{cou} = \frac{2}{M}\right)$, то сможем проинтегрировать отрезок только длины $\frac{2N}{M}$, что в N раз меньше, чем пользуясь Чебышевскими узлами.

Предложенная схема очень чувствительна к ошибкам округления. Почему? Хоть и из требования $\left\|\prod_{k=1}^{N}(1-h_kA)\right\| \le 1$, для зафиксированного k как величина h_k, так и $\|1-h_kA\|$ может быть существенно больше единицы. Действительно, пусть e - собственный нормированный вектор матрицы A, соответствующий наибольшему $\lambda(A) = M$. Найдем норму вектора $y = (I - h_NA)e$ для $i = N$:

$$h_N = \frac{1}{\frac{M}{2} + \frac{M}{2}\cos\left(\pi - \frac{\pi}{2N}\right)} = \frac{1}{\frac{M}{2} - \frac{M}{2}\cos\frac{\pi}{2N}} = \frac{2}{M\left(1 - \left(1 - \frac{1}{2}\frac{\pi^2}{4N^2} + \underline{\underline{\mathcal{O}}}(\frac{1}{N^4})\right)\right)} = \frac{2}{M\left(\frac{1}{2}\frac{\pi^2}{4N^2} + \underline{\underline{\mathcal{O}}}(\frac{1}{N^4})\right)} \approx \frac{N^2}{M}$$

Поэтому $\|y\| = \|I - h_NA\|\, e \approx \left|1 - \frac{N^2}{M}M\right| \approx N^2$. Следовательно, для больших значений N применение метода может привести как к катастрафическому росту погрешности вычислений, так и к переполнению разрядов. Эту проблему обходят перестановкой шагов таким образом, чтобы на каждом шаге $\|\prod_k(1-h_kA)\| \le 1$.

На данный момент математически обоснованы алгоритмы для $N = 2^p3^q$. Приведем без доказательства один из таких алгоритмов для $N = 2^p$. Нумеровку шагов проводят от большего к меньшему, то есть

$$h_{N+1-i} = \left(\frac{M}{2} + \frac{M}{2}\cos\frac{\pi(2i-1)}{2N}\right)^{-1},\ i = 1, \ldots, N$$

Желаемая последовательность для $N = 2$ имеет вид $(2, 1)$. Пусть построена последовательность для 2^{p-1}: $\{i_1, \ldots, i_{2^{p-1}}\}$. Тогда последовательность для 2^p имеет вид

$$\{2^p + 1 - i_1,\ i_1,\ 2^p + 1 - i_2,\ i_2, \ldots,\ 2^p + 1 - i_{2^{p-1}},\ i_{2^{p-1}}\}$$

Например: $(3, 2, 4, 1)$, $(6, 3, 7, 2, 5, 4, 8, 1)$, $(11, 6, 14, 3, 10, 7, 15, 2, 12, 5, 13, 4, 9, 8, 16, 1)$. Таким образом, делается серия по правилу "малый - большой самый малый шаг выполняется предпоследним, а самый большой последним.

18 Обыкновенные дифференциальные уравнения второго порядка, аппроксимация, α-устойчивость. Аппроксимация краевых условий третьего рода.

Обыкновенные дифференциальные уравнения второго порядка

Опр. 18.1. Будем рассматривать только задачи с правыми частями, не зависящими от y': $f(x,y,y') = f(x,y)$. вида

$$y'' = f(x,y) \tag{1}$$

Для такой задачи предлагается строить следующую разностную схему

$$\begin{gathered} h = \frac{X}{N},\ x_i = x_o + i\cdot h,\ y_h = \{y_i\}_{i=0}^N,\ \|y_h\|_{Y_h} = \max_i |y_k| \\ \begin{cases} \frac{1}{h^2}\sum_{i=0}^n a_{-i} y_{k-i} = \sum_{i=0}^n b_{-i} f_{k-i},\ k = n, n+1\dots \\ y_0, y_1, \dots, y_{n-1} - \text{ начальные условия} \end{cases} \end{gathered} \tag{2}$$

где a_{-i}, b_{-i} не зависят от h, a_0, $a_n \neq 0$ и $f_{k-i} = f(x_{k-i}, y_{k-i})$.

Хотим для такой разностной схемы проверять условие аппроксимации на решении, чтобы далее пользоваться теоремой Филиппова. Зададим функцию погрешности

$$r_h^k \overset{\text{def}}{=} \frac{1}{h^2}\sum_{i=0}^n a_{-i} y(x_{k-i}) - \sum_{i=0}^n b_{-i} f(x_{k-i}, y(x_{k-i}))$$

Условия аппроксимации на решении с порядком p на отрезке $[x_0, x_0 + X]$: $\begin{cases} \|r_h\|_{F_h} \leq ch^p \\ \|f_h - (f)_h\|_{F_h} \underset{h\to 0}{\to} 0 \end{cases}$

Коэффициенты в общем случае a_{-i} и b_{-i} находятся из условий аппроксимации на задаче

$$\|L_h(y)_{Y_h} - (Ly)_{F_h}\|_{F_h} + \|(f)_{F_h} - f_h\|_{F_h} \leq c_1 h^{p_1}$$

Теорема 18.1 (Необходимые и достаточные условия аппроксимации на решении)**.** *Для задачи* (1) *с разностной схемой* (2) *необходимые и достаточные условия аппроксимации на решении имеют вид*

$$\sum_{i=0}^n a_{-i} = 0;\ \sum_{i=0}^n b_{-i} = 1;\ \sum_{i=0}^n a_{-i} i = 0; \sum_{i=0}^n i^2 a_{-i} = 2$$

Доказательство. 1. Проверим условие нормировки правых частей

$$\|f_h - (f)_h\|_{F_h} = \left\| f(x_{k-i}) - \sum_{i=0}^n b_{-i} f(x_{k-i}) \right\|_{F_h} = \left\| f(x_{k-i}) \left(1 - \sum_{i=0}^n b_{-i}\right) \right\|_{F_h} \underset{h\to 0}{\to} 0 \Leftrightarrow \sum_{i=0}^n b_{-i} = 1$$

2. Выпишем ряд Тейлора для левой и правой части в узлах $\{x_k\}$:

$$\begin{aligned} y(x_k - ih) &= y(x_k) - ihy'(x_k) + \frac{(ih)^2}{2} y''(x_k) + \underline{\underline{\mathcal{O}}}(h^3) \\ f(x_k - ih) &= y''(x_k - ih) = y''(x_k) - ihy'''(x_k) + \underline{\underline{\mathcal{O}}}(h^2) \end{aligned}$$

Запишем условие аппроксимации на решении:

$$\begin{aligned} &\left| \frac{1}{h^2}\sum_{i=0}^n a_{-i} y_{k-i} - \sum_{i=0}^n b_{-i} f(x_k - ih) \right| \\ &= \left| \frac{1}{h^2}\sum_{i=0}^n a_{-i} y(x_k) + \frac{1}{h^2}\sum_{i=0}^n a_{-i}(-ihy'(x_k)) + \frac{1}{h^2}\sum_{i=0}^n a_{-i}\left(\frac{(ih)^2}{2} y''(x_k)\right) - \sum_{i=0}^n b_{-i} y''(x_k) + \underline{\underline{\mathcal{O}}}(h) \right| \leq \\ &\leq \left| \frac{y(x_k)}{h^2}\sum_{i=0}^n a_{-i} + \frac{y'(x_k)}{h}\left(-\sum_{i=0}^n a_{-i} i\right) + y''(x_k)\left(\sum_{i=0}^n a_{-i}\frac{i^2}{2} - \sum_{i=0}^n b_{-i}\right) + \underline{\underline{\mathcal{O}}}(h) \right| \leq ch \end{aligned}$$

Для выполнения условия аппроксимации нужно, чтобы все коэффициенты до h занулились, отсюда и из проверки нормировки следует доказательство теоремы.

□

Замечание 18.1. Для того, чтобы найти a_{-i} и b_{-i} надо решить систему уравнений. Для того, чтобы система была разрешима, важно проверить $p = 2n$.

α-устойчивость задачи второго порядка

В случае задачи (1) проверяют более слабое определение α-устойчивости.

Опр. 18.2. Для задачи Коши $y'' = f(x)$ схема называется α- устойчивой, если все корни соответствующего характеристического многочлена $\sum_{i=0}^{n} a_{-i}\mu^{k-i}$ однородного уравнения принадлежат единичному кругу и на границе круга нет кратных корней, за исключением $\mu = 1$ кратности 2.

Отличие условия устойчивости для задачи Коши второго порядка от условия устойчивости для задачи первого порядка обусловлено более высокой степенью h в правой части разностной схемы $\sum_{i=0}^{n} a_{-i}y_{k-i} = h^2 \sum_{i=0}^{n} b_{-i}f_{k-i}$.

Аппроксимация граничных условий третьего рода

Пусть дана задача

$$\begin{cases} -(k(x)y')' + p(x)y = f(x),\ 0 < k_0 < k(x) < k_1,\ 0 \le p(x) \le p_1 \\ ay + by' = c \end{cases}$$

Предлагается выбрать следующую разностную схему

$$-\frac{1}{h}\left[k(x_{i+1/2})\frac{y_{i+1} - y_i}{h} - k(x_{i-1/2})\frac{y_i - y_{i-1}}{h}\right] + p(x_i)y_i = f_i$$

Проверим, что она второго порядка аппроксимации на решении:

$$\left| -\frac{1}{h}\underbrace{\left[k\left(x_k + \frac{h}{2}\right)\frac{y(x_k + h) - y(x_k)}{h} - k\left(x_k - \frac{h}{2}\right)\frac{y(x_k) - y(x_k - h)}{h}\right]}_{\star} + p(x_k)y(x_k) - f(x_k)\right| \le ch^2$$

Напомним формулу разложения в ряд Тейлора в точке x_k

$$u(x \pm h) = u(x) \pm hu'(x) + \frac{h^2}{2}u''(x) + \underline{\underline{\mathcal{O}}}(h^3)$$

Отдельно решим то, что помечено $\star$:

$$\left(k(x_k) + \frac{h}{2}k'(x_k) + \frac{h^2}{8}k''(x_k) + \underline{\underline{\mathcal{O}}}(h^3)\right)\left(y'(x_k) + \frac{h}{2}k''(x_k) + \underline{\underline{\mathcal{O}}}(h^2)\right) -$$
$$\left(k(x_k) - \frac{h}{2}k'(x_k) + \frac{h^2}{8}k''(x_k) + \underline{\underline{\mathcal{O}}}(h^3)\right)\left(y'(x_k) - \frac{h}{2}k''(x_k) + \underline{\underline{\mathcal{O}}}(h^2)\right) =$$
$$= hk'(x_k)y'(x_k) + hk(x_k)y''(x_k) + \underline{\underline{\mathcal{O}}}(h^3) = h(k(x_k)y'(x_k))' + \underline{\underline{\mathcal{O}}}(h^3)$$

Подставим получившееся значение и начальное условие в изначальное уравнение:

$$\left|-(k(x_k)y'(x_k))' + \underline{\underline{\mathcal{O}}}(h^2) + p(x_k)y(x_k) - (-(k(x_k)y'(x_k))' + p(x_k)y(x_k))\right| \le ch^2$$

Действительно, предложенная схема обладает вторым порядком аппроксимации на решении.

Построим для краевого условия задачи - краевого условия *третьего рода* - конечно-разностную аппроксимацию второго порядка точности на решении, используя значения функции y в точках $x_0 = 0$ и $x_1 = h$. Для простоты возьмем $k(x) \equiv 1$. Воспользуемся δ-поправкой.

Хотим получить следующее:

$$\left|ay(0) + b\frac{y(h) - y(0)}{h} - c - \delta\right| \le ch^2$$

Из формулы Тейлора в точке 0 имеем:

$$\left|ay(0) + b\frac{y(0) + hy'(0) + \frac{h^2}{2}y''(0) + \underline{\underline{\mathcal{O}}}(h^3) - y(0)}{h} - c - \delta\right| \le ch^2$$

$$\left|ay(0) + by'(0) + \frac{bh}{2}y''(0) + \underline{\underline{\mathcal{O}}}(h^2) - c - \delta\right| \le ch^2$$

$$\left|ay(0) + by'(0) + \frac{bh}{2}(p(0)y(0) - f(0)) + \underline{\underline{\mathcal{O}}}(h^2) - c - \delta\right| \le ch^2$$

Чтобы достигалось соответствующее неравенство требуется взять $\delta := \frac{bh}{2}(p(0)y(0) - f(0))$.

Таким образом аппроксимация второго порядка на краевом условии имеет вид

$$ay_0 + b\frac{y_1 - y_0}{h} = c + \frac{bh}{2}(p(0)y_0 - f(0))$$

Примеры

Рассмотрим разностные схемы для уравнения $y''(x) = f(x)$

Пример 18.1. Естественная аппроксимация:

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} = f_k$$

Главный член погрешности на решении равен

$$r_h := L_h(y)_h - f_h = \frac{h^2}{12} y^{(4)}(x_k) + \underline{\underline{\mathcal{O}}}(h^4)$$

Пример 18.2. Схема

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} = f_k \frac{h^2}{12} f''(x_k)$$

аппроксимирует уравнение на решении с порядком $\underline{\underline{\mathcal{O}}}(h^4)$ (см. предыдущий пример и пользуемся тем, что $y'' = f$).

Пример 18.3. Схема Нумерова

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} = \frac{f_{k+1} + 10f_k + f_{k-1}}{12}$$

Главный член погрешности имеет вид

$$r_h := L_h(y)_h - f_h = \frac{h^2}{12} y^{(4)}(x_k) + \frac{h^4}{360} y^{(6)}(x_k) - \frac{h^2}{12} f^{(2)}(x_k) - \frac{h^4}{144} f^{(4)}(x_k) = -\frac{h^4}{240} y^{(6)}(x_k) + \underline{\underline{\mathcal{O}}}(h^6)$$

Такую схему используют, если невозможно посчитать f''.

19 Устойчивость краевой задачи для уравнения второго порядка: метод собственных функций.

Напомним определение устойчивости

Опр. 19.1. Разностная схема $\begin{cases} L_h y_h = f_h \\ l_h y_h = \varphi_h \end{cases}$ называется устойчивой, если: $\forall y_h^{(1)},\ y_h^{(2)}\ \forall \varepsilon > 0\ \exists \delta = \delta(\varepsilon)$:

$$\left\| f_h^{(1)} - f_h^{(2)} \right\| + \left\| \varphi_h^{(1)} - \varphi_h^{(2)} \right\| \le \delta,\ \forall h \le h_0 \Rightarrow \left\| y_h^{(1)} - y_h^{(2)} \right\| \le \varepsilon$$

Опр. 19.2. Линейная схема $\begin{cases} L_h y_h = f_h \\ l_h y_h = \varphi_h \end{cases}$ называется устойчивой, если:

$$\left\| y_h^{(1)} - y_h^{(2)} \right\| \le C \left(\left\| f_h^{(1)} - f_h^{(2)} \right\| + \left\| \varphi_h^{(1)} - \varphi_h^{(2)} \right\| \right),\ \forall h \le h_0$$

C не должна зависеть от h.

Если в разностной схеме матрица L_h является не вырожденной, то $y_h = L_h^{-1} f_h$. Отсюда следует неравенство для нормы векторов

$$\left\| y_h^{(1)} - y_h^{(2)} \right\|_h \le \left\| L_h^{-1} \right\|_h \left\| f_h^{(1)} - f_h^{(2)} \right\|_h$$

То есть, чтобы удостовериться в том, что схема устойчива надо выбрать $C \ge \left\| L_h^{-1} \right\|_h$.

Исследуем устойчивость разностной схемы

$$-\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} = f_k,\ y_0 = y_N = 0,\ h = 1/N \Leftrightarrow L_h y_h = f_h \tag{1}$$

в сеточной интегральной норме $\|y_h\|_h^2 = (y_h, y_h)_h = \sum_{k=1}^{N-1} y_k^2$, согласованной с непрерывной нормой $\|y(x)\|_{L_2(0,1)}^2 = \int_0^1 y^2(x)dx$ исходной задачи. Будем оценивать норму оператора $\left\| L_h^{-1} \right\|_h$, где

$$\left\| L_h^{-1} \right\|_h^2 \overset{\text{def}}{=} \sup_{y_h \ne 0} \frac{\left\| L_h^{-1} y_h \right\|_h}{\|y_h\|_h} = \sup_{y_h \ne 0} \frac{(L_h^{-1} y_h, L_h^{-1} y_h)_h}{(y_h, y_h)_h}$$

Пусть известны собственные числа λ_n матрицы L_h, собственные вектора ортонормальны: $(y^{(n)}, y^{(m)}) = \delta_m^n$. Тогда

$$\left\| L_h^{-1} \right\|_h^2 = \sup_{\{c_n\}} \frac{(\sum \lambda_n^{-1} c_n y_h^{(n)}, \sum \lambda_n^{-1} c_n y_h^{(n)})_h}{(\sum c_n y_h^{(n)}, \sum c_n y_h^{(n)})_h} = \sup_{\{c_n\}} \frac{\sum \lambda_n^{-2} c_n^2}{\sum c_n^2} = \max_n \left| \lambda_n^{-2} \right| \sup_{\{c_n\}} \frac{\sum c_n^2}{\sum c_n^2} = \max_n \left| \lambda_n^{-2} \right|$$

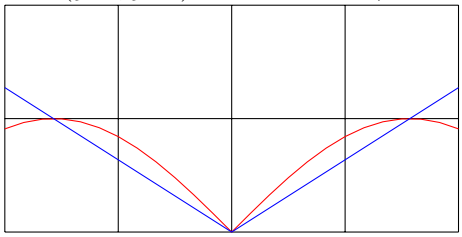
То есть $\|L_h\|_h = \lambda_{\min}^{-1}$. Решив разностную задачу (1) аналитически, получим

$$y_k^{(n)} = \sin \pi n k h,\ \lambda_n = \frac{4}{h^2} \sin^2 \frac{\pi n h}{2},\ n = 1, \dots, N-1$$

Проверим, что получившиеся решения уравнения являются ортогональными векторами. Матрица L_h симметрична $L_h = L_h^T$, а это значит, что

$$(L_h y^{(n)}, y^{(m)})_h = (y^{(n)}, L_h y^{(m)})_h \Leftrightarrow 0 = (L_h y^{(n)}, y^{(m)})_h - (y^{(n)}, L_h y^{(m)})_h = (\lambda^{(n)} - \lambda^{(m)})(y^{(n)}, y^{(m)})_h$$

То есть $(y^{(n)}, y^{(m)})_h = 0$ для $\lambda^{(m)} \ne \lambda^{(n)}$. Ортогональность собственных векторов доказана.



Из неравенства $\sin|x| \ge \frac{2}{\pi}|x|$ при $|x| \le \frac{\pi}{2}$ имеем

$$\lambda_{\min} = \lambda_1 = \frac{4}{h^2} \sin^2 \frac{\pi h}{2} \ge \frac{4}{h^2} h^2 \ge 4$$

$$\lambda_{\max} = \lambda_{N-1} = \frac{4}{h^2} \sin^2 \frac{\pi h (N-1)}{2} \le \frac{4}{h^2} \sin^2 \frac{\pi}{2} \le \frac{4}{h^2}$$

Таким образом, верна не зависящая от h оценка для нормы $\left\| L_h^{-1} \right\|_h \le \frac{1}{4}$, то есть мы доказали устойчивость схемы по определению.

20 Устойчивость краевой задачи для уравнения второго порядка: энергетический метод.

Напомним определение устойчивости

Опр. 20.1. Разностная схема $\begin{cases} L_h y_h = f_h \\ l_h y_h = \varphi_h \end{cases}$ называется устойчивой, если: $\forall y_h^{(1)},\ y_h^{(2)}\ \forall \varepsilon > 0\ \exists \delta = \delta(\varepsilon)$:

$$\left\| f_h^{(1)} - f_h^{(2)} \right\| + \left\| \varphi_h^{(1)} - \varphi_h^{(2)} \right\| \le \delta,\ \forall h \le h_0 \Rightarrow \left\| y_h^{(1)} - y_h^{(2)} \right\| \le \varepsilon$$

Опр. 20.2. Линейная схема $\begin{cases} L_h y_h = f_h \\ l_h y_h = \varphi_h \end{cases}$ называется устойчивой, если:

$$\left\| y_h^{(1)} - y_h^{(2)} \right\| \le C \left(\left\| f_h^{(1)} - f_h^{(2)} \right\| + \left\| \varphi_h^{(1)} - \varphi_h^{(2)} \right\| \right),\ \forall h \le h_0$$

C не должна зависеть от h.

Будем доказывать устойчивость разностной схемы энергетическим методом. Запишем нашу дифференциалную задачу

$$-y''(x) + p(x)y(x) = f(x),\ y(0) = y'(1) = 0,\ p(x) \ge 0$$

Умножим уравнение на $y(x)$, и результат проинтегрируем по отрезку $[0, 1]$

$$\int_0^1 (-y''y + py^2)dx = \int_0^1 fydx$$

$$\int_0^1 -y''ydx + \int_0^1 py^2dx = \int_0^1 fydx$$

Проинтегрируем по частям первое слагаемое

$$\int_0^1 -y''ydx = \int_0^1 -ydy' = -yy'|_0^1 - \int_0^1 y'd(-y) = \int_0^1 (y')^2dx$$

Получили интегральное тождество

$$\int_0^1 (y'(x))^2dx + \int_0^1 py^2dx = \int_0^1 fydx$$

Оценим слева через неравенство, связывающее интегралы от квадратов функции и ее производной. Так как $y(0) = 0$, то справедливо следующее:

$$y(x_0) = \int_0^{x_0} y'(x)dx$$

Применим интегральную форму неравенства Коши-Буняковского:

$$|y(x_0)|^2 = \left| \int_0^{x_0} y'dx \right|^2 \le \left(\int_0^{x_0} 1^2dx \right) \left(\int_0^{x_0} (y')^2dx \right) \le \int_0^{x_0} (y')^2dx \le \int_0^1 (y')^2dx$$

После интегрирования по x_0 по отрезку $[0, 1]$ обеих частей получим искомое равенство

$$\int_0^1 |y(x_0)|^2dx_0 \le \int_0^1 (y')^2dx \int_0^1 dx_0 \Leftrightarrow \int_0^1 y^2dx \le \int_0^1 (y')^2dx$$

Оценку справа выведем из разности квадратов:

$$0 \le \int_0^1 (f - y)^2dx \le \int_0^1 f^2dx - 2\int_0^1 fydx + \int_0^1 y^2dx$$

$$\Rightarrow \int_0^1 fydx \le \frac{1}{2} \left(\int_0^1 f^2dx + \int_0^1 y^2dx \right)$$

Таким образом, имеем:

$$\int_0^1 y^2 dx \le \int_0^1 (y'(x))^2 dx + \int_0^1 py^2 dx = \int_0^1 fy dx \le \frac{1}{2}\left(\int_0^1 f^2 dx + \int_0^1 y^2 dx\right)$$

Получаем следующую оценку

$$\int_0^1 y^2 dx \le \int_0^1 f^2 dx \Rightarrow \|y\|_{L_2(0,1)} \le \|f\|_{L_2(0,1)}$$

Это означает устойчивость дифференциальной задачи по правой части.

Докажем теперь устойчивость разностной схемы.

$$-\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} + p_k y_k = f_k,\ 1 \le k \le N-1,\ y_0 = 0,\ y_N = y_{N-1}$$

Умножим на y_k и просуммируем от 1 до $N-1$. Так как $y_0 = 0,\ y_N = y_{N-1}$

$$-\frac{1}{h^2}\left(\sum_{k=1}^{N-1}(y_{k+1} - 2y_k + y_{k-1})\, y_k\right) = -\frac{1}{h^2}\left(\sum_{k=1}^{N-1}(y_{k+1} - y_k - y_k + y_{k-1})\, y_k\right) =$$

$$= -\frac{1}{h^2}\sum_{k=1}^{N-1}(y_{k+1} - y_k)\, y_k + \frac{1}{h^2}\sum_{k=1}^{N-1}(y_k - y_{k-1})\, y_k = -\frac{1}{h^2}\sum_{k=2}^{N}(y_k - y_{k-1})\, y_{k-1} + \frac{1}{h^2}\sum_{k=1}^{N-1}(y_k - y_{k-1})\, y_k =$$

$$= -\frac{1}{h^2}\sum_{k=2}^{N}\left(-(y_k - y_{k-1})\, y_{k-1} + (y_k - y_{k-1})\, y_k\right) = \frac{1}{h^2}\sum_{k=1}^{N}(y_k - y_{k-1})^2$$

Получили конечномерный аналог интегрального тождества:

$$\frac{1}{h^2}\sum_{k=1}^{N}(y_k - y_{k-1})^2 + \sum_{k=1}^{N-1} p_k y_k^2 = \sum_{k=1}^{N-1} f_k y_k$$

Для оценки слева докажем сеточный аналог неравенства для функции и ее производной в точках $k = 1, ..., N-1$. Так как $y_0 = 0$, справедливо следующее:

$$y_k = \sum_{i=1}^{k}(y_i - y_{i-1})$$

Воспользуемся неравенством Коши-Буняковского и $y_N = y_{N-1}$

$$y_k^2 \le \left(\sum_{i=1}^{k} 1^2\right)\left(\sum_{i=1}^{k}(y_i - y_{i-1})^2\right) \le (N-1)\sum_{i=1}^{N-1}(y_i - y_{i-1})^2$$

Суммируя до $N-1$ обе части, при $h = \frac{2}{2N-1}$ получаем оценку:

$$\sum_{k=1}^{N-1} y_k^2 \le (N-1)^2 \sum_{k=1}^{N-1}(y_k - y_{k-1})^2 \le \frac{1}{h^2}\sum_{k=1}^{N-1}(y_k - y_{k-1})^2$$

Найдем аналогично дифференциальному неравенству оценку справа

$$0 \le \sum_{k=1}^{N-1}(f_k - y_k)^2 = \sum_{k=1}^{N-1} f_k^2 - 2\sum_{k=1}^{N-1} f_k y_k + \sum_{k=1}^{N-1} y_k^2$$

$$\Rightarrow \sum_{k=1}^{N-1} f_k y_k \le \frac{1}{2}\left(\sum_{k=1}^{N-1} f_k^2 + \sum_{k=1}^{N-1} y_k^2\right)$$

Итоговая оценка имеет вид

$$\sum_{k=1}^{N-1} y_k^2 \le \frac{1}{h^2}\sum_{k=1}^{N-1}(y_k - y_{k-1})^2 + \sum_{k=1}^{N-1} p_k y_k^2 = \sum_{k=1}^{N-1} f_k y_k \le \frac{1}{2}\left(\sum_{k=1}^{N-1} f_k^2 + \sum_{k=1}^{N-1} y_k^2\right)$$

Таким образом,

$$\sum_{k=1}^{N-1} y_k^2 \le \sum_{k=1}^{N-1} f_k^2 \Rightarrow \sum_{k=1}^{N-1} y_k^2 h \le \sum_{k=1}^{N-1} f_k^2 h \Rightarrow \|y_h\|_h^2 \le \|f_h\|_h^2$$

То есть **разностная схема устойчива** в норме $\|\cdot\|_h$.

21 Метод прогонки.

Требуется найти решение y задачи $Ay = f,\ A \in \mathbb{R}^{N-1\times N-1}$ с заданной матрицей квадратной матрицей A (примеры см. ниже), заданной правой частью f.

Перепишем матрицу A, сделав вспомогательные замены:

$$\begin{pmatrix} c_1 & -b_1 & & & & 0 \\ -a_2 & c_2 & -b_2 & & & \\ & \cdots & \cdots & & & \\ & & -a_{N-2} & c_{N-2} & -b_{N-2} & \\ 0 & & & -a_{N-1} & c_{N-1} & \end{pmatrix}$$

Тогда задачу можно переписать следующем образом

$$\begin{array}{cc} c_1y_1 - b_1y_2 = f_1, & k = 1 \\ -a_ky_{k-1} + c_ky_k - b_ky_{k+1} = f_k & k = 2,\dots,N-2 \\ -a_{N-1}y_{N-2} + c_{N-1}y_{N-1} = f_{N-1}, & k = N-1 \end{array} \tag{1}$$

Перепишем первое уравнение

$$c_1y_1 - b_1y_2 = f_1 \Leftrightarrow y_1 - \frac{b_1}{c_1}y_2 = \frac{f_1}{c_1} \Leftrightarrow y_1 = \alpha_2y_2 + \beta_2,\ \alpha_2 = \frac{b_1}{c_1},\ \beta_2 = \frac{f_1}{c_1}$$

Найдем α_{k+1} и β_{k+1}, используя полученную формулу $y_{k-1} = \alpha_ky_k + \beta_k$, подставив во второе уравнение

$$-a_k(\alpha_ky_k + \beta_k) + c_ky_k - b_ky_{k+1} = f_k \Leftrightarrow (-a_k\alpha_k + c_k)y_k - a_k\beta_k - b_ky_{k+1} = f_k \Leftrightarrow$$

$$\Leftrightarrow (-\alpha_ka_k + c_k)y_k + (-b_k)y_{k+1} = a_k\beta_k + f_k \Leftrightarrow y_k = \left(\frac{b_k}{c_k - \alpha_ka_k}\right)y_{k+1} + \frac{a_k\beta_k + f_k}{c_k - \alpha_ka_k}$$

$$\alpha_{k+1} = \frac{b_k}{c_k - \alpha_ka_k},\ \beta_{k+1} = \frac{a_k\beta_k + f_k}{c_k - \alpha_ka_k}$$

Рассмотрим последнее равенство, подставим в него $y_{N-2} = \alpha_{N-1}y_{N-1} + \beta_{N-1}$

$$-a_{N-1}(\alpha_{N-1}y_{N-1} + \beta_{N-1}) + c_{N-1}y_{N-1} = f_{N-1}$$

$$y_{N-1} = \frac{f_{N-1} + a_{N-1}\beta_{N-1}}{c_{N-1} - a_{N-1}\alpha_{N-1}};\ y_k = \alpha_{k+1}y_{k+1} + \beta_{k+1},\ k = N-2,\dots,1$$

Получили формулы для *правой прогонки.*

Теорема 21.1 (Достаточные условия корректности и устойчивости метода прогонки)**.** *Пусть коэффициенты* (1) *действительны и удовлетворяют условиям:* c_1, c_{N-1}, a_k, c_k, b_k *при* $k = 2,\dots,N-2$ *отличны от нуля и*

$$|c_k| \geq |a_k| + |b_k|,\ k = 2,\dots,N-2$$

$$|c_1| \geq |b_1|;\ |c_{N-1}| \geq |a_{N-1}|$$

При чем хотя бы одно из неравенств является строгим. Тогда для формул метода прогонки справедливы неравенства

$$c_k - a_k\alpha_k \neq 0,\ |a_k| \leq 1,\ k = 2,\dots,N-1$$

гарантирующие разрешимость и устойчивость, то есть корректность метода.

Доказательство. Убедимся методом индукции, что ни один из знаменателей не обращается в ноль, то есть убедимся в устойчивости метода. База: $\left|\frac{b_1}{c_1}\right| = |\alpha_2| \leq 1$ из условий теоремы. Шаг: Пусть $|\alpha_k| \leq 1$.

$$|c_k - a_k\alpha_k| \geq |c_k| - |a_k|\,|\alpha_k| \geq |c_k| - |\alpha_k| \geq |b_k| > 0$$

$$\Rightarrow |\alpha_{k+1}| = \frac{|b_k|}{|c_k - \alpha_ka_k|} \leq 1$$

Обратим внимание, что если $|\alpha_{k_0}| < 1$, то $\forall k > k_0\ \alpha_k < 1$.

Проверим, что последнее слагаемое так же не обратится в ноль. Рассмотрим

$$|c_{N-1} - \alpha_{N-1}a_{N-1}| \geq |c_{N-1}| - |\alpha_{N-1}|\,|a_{N-1}|$$

Здесь нам потребуется условие, что хотя бы одно из неравенств должно обращаться в строгое равенство.

1. Если $|c_{N-1}| > |a_{N-1}|$, то $|c_{N-1}| - |\alpha_{N-1}|\,|a_{N-1}| > 0$, значит и $|c_{N-1} - \alpha_{N-1}a_{N-1}| > 0$, ч.т.д.
2. Если $|c_k| > |a_k| + |b_k|$, то $|c_k| - |a_k| > |b_k| > 0$, значит и $|c_{N-1} - \alpha_{N-1}a_{N-1}| > 0$, ч.т.д.
3. Если $|c_1| > |b_1|$, то $|\alpha_2| < 1 \Rightarrow |\alpha_{N-1} < 1|$, значит и $|c_{N-1} - \alpha_{N-1}a_{N-1}| > 0$, ч.т.д.

□

Пример 21.1. Для задачи $-y'' = f,\ y(0) = a,\ y(1) = b$ разностная аппроксимация имеет вид

$$\frac{1}{h^2}\underbrace{\begin{pmatrix} 2 & -1 & & & & 0 \\ -1 & 2 & -1 & & & \\ & \cdots & \cdots & & & \\ & & & -1 & 2 & -1 \\ 0 & & & & -1 & 1 \end{pmatrix}}_{N-1\times N-1}\begin{pmatrix} y_1 \\ \vdots \\ y_{N-1} \end{pmatrix} = \begin{pmatrix} f_1 + \frac{a}{h^2} \\ \vdots \\ f_{N-1} + \frac{b}{h^2} \end{pmatrix}$$

В данном примере мы исключили известные значениея y_0, y_N. Метод прогонки устойчив.

Пример 21.2. Для задачи $-y'' = f,\ y'(0) = a,\ y'(1) = b$ разностная аппроксимация имеет вид

$$\frac{1}{h^2}\underbrace{\begin{pmatrix} 2 & -2 & & & & 0 \\ -1 & 2 & -1 & & & \\ & \cdots & \cdots & & & \\ & & & -1 & 2 & -1 \\ 0 & & & & -2 & 2 \end{pmatrix}}_{N+1\times N+1}\begin{pmatrix} y_0 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} f_0 - \frac{2a}{h} \\ \vdots \\ f_N - \frac{2b}{h} \end{pmatrix}$$

Аппроксимация для краевых условий здесь подобрана с помощью δ-поправки. Метод прогонки в данной задаче не устойчив, так как не выполняется условие о наличии хотя бы одного строгого неравенства.

22 Метод стрельбы и метод Фурье. Численные методы линейной алгебры.

Метод стрельбы

Решаем следующую конечно-разностную схему

$$-\frac{y_{k+1}-2y_k+y_{k-1}}{h^2}=f,\ y_0=a,\ y_N=b \Leftrightarrow L_h y_h=f_h$$

Если бы у нас было начальное условие $y_1=c$, то мы могли бы однозначно построить решение, используя методы решения разностных уравнений, но у нас такого нет. Идея этого метода следующая: рассмотрим две вспомогательные задачи

$$L_h u_h=f_h,\ u_0=a,\ u_1=\hat{u}$$

$$L_h w_h=0,\ w_0=0,\ u_1=\hat{w}$$

Такие две задачи мы можем решить рекуррентно пересчитывая каждое следующее значение. Итоговое решение предлагается искать в виде

$$y_h=u_h+Cw_h \tag{1}$$

где C мы получим из второго начального условия изначальной задачи. Почему это верно? Применим разностный оператор к обеим частям:

$$L_h y_h=L_h u_h+C\underbrace{L_h w_h}_{=0}=f_h$$

Ищем C следующим образом: рекуррентно посчитаем u_N и w_N и подставим в (1): $C=\frac{y_N-u_N}{w_N}=\frac{b-u_N}{w_N}$

Формально значения $\hat{u}$ и $\hat{w}$ могут быть произвольными, но разумно выбирать их так, чтобы $u_1-u_0=O(h)$, $w_1-w_0=O(h)$. Потому что переход от u_0 к u_1 идет за один шаг (h), а если выбрать очень их большими, то итоговый ответ будет испорчен из-за погрешности вычислений.

Замечание 22.1. Название метода берет свое начало из аналогичного метода прицеливания артиллерией: сначала наводятся на цель, используя одни начальные данные, и смотрят на результирующую траекторию снаряда, затем выбирают другие начальные данные, получают другую траекторию, и в конце высчитывают те начальные данные, которые точно гарантируют попадание.

Метод Фурье

Требуется найти решение СЛАУ $Ay=f,\ y,f\in\mathbb{R},\ A\in\mathbb{R}^{M\times M}$. Известны собственные вектора и собственные числа матрицы $Ay^{(n)}=\lambda_n y^{(n)},\ n=1,\dots,M$ и $y^{(n)}$ образуют ортонормированный базис в R^M.

Идея метода состоит в разложении решения y по базисным векторам $y=\sum_{n=1}^{M}c_n y^{(n)}$, определении коэффициентов c_n и последующем восстановлении y.

Замечание 22.2. Вспомним, что нахождение определителя матрицы является неустойчивой задачей, и, соответственно, обычный способ нахождения собственных значений через характеристический многочлен нам не подходит. Существуют и другие способы их нахождения, но сложность таких алгоритмов не меньше использования алгоритма с нахождением определителя. Поэтому метод Фурье чаще используют тогда, когда все собственные числа и вектора возможно найти аналитически, и система собственных векторов образует ортонормированный базис в пространстве решений относительно зафиксированного скалярного произведения $(y^{(m)},y^{(n)})_h=\delta_m^n$. Когда все эти условия соблюдаются, то коэффициенты c_n могут быть найдены по явной формуле.

Перепишем исходную задачу, выразив решение через базис

$$A\left(\sum_{n=1}^{M}c_n y^{(n)}\right)=b\Leftrightarrow\sum_{n=1}^{M}\lambda_n c_n y^{(m)}=b$$

Пройдемся по всем $m=1,\dots,M$, скалярное умножим данное равенство на $y^{(m)}$ и воспользуемся ортонормированностью базиса

$$\left(\sum_{n=1}^{M}\lambda_n c_n y^{(n)},y^{(m)}\right)_h=(b,y^{(m)})_h\Leftrightarrow\lambda_m c_m=(b,y^{(m)})_h$$

Рассмотрим правую часть: аналогично разложим b по базису и получим $\left(\sum_{m=1}^{M} d_m y^{(m)} y, y^{(m)}\right)_h = d_m$. Отсюда имеем $c_m = \frac{d_m}{\lambda_m}$. Итоговый ответ имеет вид

$$y = \sum_{m=1}^{M} \frac{d_m}{\lambda_m} y^{(m)}$$

Замечание 22.3. Обратим внимание, что в получившемся решении $\lambda_m \neq 0$. Иначе если $\lambda_m = 0$, то $\det A = 0$, а это не может нам гарантировать корректность исходной задачи, то есть решение исходной задачи не является единственным.

Замечание 22.4. По сравнению с методом прогонки, у которого сложность $O(N)$, у метода Фурье в общем случае $\underline{\underline{\mathcal{O}}}(N^2)$. Эту сложность можно уменьшить за счет так называемого быстрого преобразования Фурье ($\underline{\underline{\mathcal{O}}}(N \log N)$).

Пример 22.1. Решим задачу $-y'' + py = f,\ p \equiv \text{const} \geq 0,\ y(0) = y(1) = 0$ методом Фурье. Воспользуемся следующей конечно-разностной схемой

$$-\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} + py_k = f_k,\ y_0 = y_N = 0,\ k = 1, \ldots, N-1$$

Выберем согласованное скалярное произведение: $(u, v)_h = h \sum_{i=1}^{N-1} u_i v_i$.

Матричная запись данной разностной схемы имеет вид

$$\left[-\begin{pmatrix} \frac{-2}{h^2} & \frac{1}{h^2} & & & & 0 \\ \frac{1}{h^2} & \frac{-2}{h^2} & \frac{1}{h^2} & & & \\ & \cdots & \cdots & \cdots & & \\ & & & \frac{1}{h^2} & \frac{-2}{h^2} & \frac{1}{h^2} \\ 0 & & & & \frac{1}{h^2} & \frac{-1}{h^2} \end{pmatrix} + \begin{pmatrix} p_1 & & 0 \\ & \ddots & \\ 0 & & p_{N-1} \end{pmatrix} \right] \begin{pmatrix} y_1 \\ \vdots \\ y_{N-1} \end{pmatrix} = \begin{pmatrix} f_1 \\ \vdots \\ f_{N-1} \end{pmatrix} \Leftrightarrow \left[\tilde{A} + pI\right] y = f$$

Решение данной разностной задачи может быть найдено аналитически

$$y_k^{(m)} = C \sin(\pi k m h),\ \lambda_m = \frac{4}{h^2} \sin^2\left(\frac{\pi m h}{2}\right) + p,\ k = 0, \ldots, N,\ m = 1, \ldots, N-1$$

Замечание 22.5. Обратим внимание почему $p \equiv \text{const} \geq 0$

$$\tilde{A}\tilde{e_i} = \tilde{\lambda_i}\tilde{e_i} \Rightarrow (A - pI)\tilde{e_i} = A\tilde{e_i} - p\tilde{e_i} = \tilde{\lambda_i}\tilde{e_i} \Leftrightarrow A\tilde{e_i} = (\tilde{\lambda_i} + p)\tilde{e_i}$$

Мы можем ослабить условие на p, сохраняя необходимое условие метода Фурье: $\lambda_m = \frac{4}{h^2} \sin^2\left(\frac{\pi m h}{2}\right) + p \neq 0$

Проверим, что для полученного решения выполняются необходимые условия для применения метода Фурье: ортогональность векторов $y^{(m)}$ гарантируется так как исходная матрица является симметричной. Подберем C так, чтобы выполнялась ортонормированность:

$$(y^{(n)}, y^{(n)})_h = h \sum_{k=1}^{N-1} (y_k^{(n)})^2 = hC^2 \sum_{k=1}^{N-1} \sin^2(\pi k n h) = \frac{hC^2}{2} \sum_{k=1}^{N-1} (1 - \cos(2\pi k n h)) = \frac{hC^2}{2}(N-1) - \frac{hC^2}{2} \cdot \star$$

$$\star : \sum_{k=1}^{N-1} \cos(2\pi k n h) = \operatorname{Re} \sum_{k=1}^{N-1} e^{2\pi i k n h} = \operatorname{Re} \left(\underbrace{\sum_{k=1}^{N} e^{2\pi i k n h}}_{=0} - e^{2\pi i n N \frac{1}{N}} \right) = -\operatorname{Re}(e^{2\pi i n}) = -1$$

$$(y^{(n)}, y^{(n)})_h = \frac{hC^2}{2}(N-1) + \frac{hC^2}{2} = 1 \Leftrightarrow \frac{NhC^2}{2} = 1 \Leftrightarrow C = \sqrt{2}$$

Таким образом, при $C = \sqrt{2}$ гарантируем ортонормированность собственных векторов. Можем применить метод Фурье.

Пример 22.2. Решим задачу $-y'' = f,\ y'(0) = y'(1) = 0$ методом Фурье. Воспользуемся следующей конечно-разностной схемой

$$\begin{cases} -\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} = f_k,\ k = 1, \ldots, N-1,\ h = 1/N \\ \frac{2}{h^2}(y_1 - y_0) = f_0 \\ \frac{2}{h^2}(y_N - y_{N-1}) = f_N \end{cases}$$

Выберем согласованное скалярное произведение: $(u,v)_h = h\sum_{i=1}^{N-1} u_i v_i + \frac{h}{2}u_0 v_0 + \frac{h}{2}u_N v_N$.
Матричная запись данной разностной схемы имеет вид

$$-\begin{pmatrix} \frac{-2}{h^2} & \frac{2}{h^2} & & & & 0 \\ \frac{1}{h^2} & \frac{-2}{h^2} & \frac{1}{h^2} & & & \\ & & \cdots & \cdots & & \\ & & & \frac{1}{h^2} & \frac{-2}{h^2} & \frac{1}{h^2} \\ 0 & & & & \frac{2}{h^2} & \frac{-2}{h^2} \end{pmatrix} \begin{pmatrix} y_0 \\ \\ \vdots \\ \\ y_N \end{pmatrix} = \begin{pmatrix} f_0 \\ \\ \vdots \\ \\ f_N \end{pmatrix}$$

Решение данной разностной задачи может быть найдено аналитически

$$y_k^{(m)} = C\cos(\pi k m h),\ \lambda_m = \frac{4}{h^2}\sin^2\left(\frac{\pi m h}{2}\right),\ k = 0,\dots,N,\ m = 0,\dots,N$$

Обратим внимание, что матрица не является симметричной в метрике исходной задачи. Покажем, что в выбранной метрике матрица как оператор, действующий на вектор y, является симметричной:

$$(Au,v)_h = h\sum_{i=1}^{N-1}\sum_{j=0}^{N} a_{ij}u_i v_i + \frac{h}{2}\sum_{j=0}^{N} a_{0j}u_0 v_0 + \frac{h}{2}\sum_{j=0}^{N} a_{Nj}u_N v_N = h(\tilde{A}u,v),\ \frac{1}{2}a_{0j} = \tilde{a}_{0j},\ \frac{1}{2}a_{Nj} = \tilde{a}_{Nj}$$

Но матрица $\tilde{A} = \tilde{A}^T$, следовательно:

$$(Au,v)_h = h(\tilde{A}u,v) = h(u,\tilde{A}v) = h(\tilde{A}v,u) = (Av,u)_h = (u,Av)_h$$

Проделав шаги, аналогичные предыдущему примеру, для соблюдения ортонормированности получим соответствующие константы $C_k = \sqrt{2},\ k = 1,\dots,N-1$, $C_0 = C_N = 1$.

Отметим, что $\lambda_0 = 0$, $y^{(0)} \equiv 1$. Но для применимости метода Фурье требуется, чтобы собственные значения были не нулевые. Вспомним для чего это требуется: при поиске коэффициентов c_m в разложении решения в базисе собственных векторов мы пришли к равенству $\lambda_m c_m = (b, y^{(m)})_h$. Тогда необходимым и достаточным условием корректности алгоритма требуется, чтобы $(b,1)_h = 0$.

23 Нормы векторов, линейных операторов, обусловленность матрицы. Оценка возмущения решения системы линейных алгебраических уравнений при возмущении правой части.

Опр. 23.1. Нормой вектора $\mathbf{x} = (x_1, \dots, x_n)^\intercal$ называется функционал, обозначаемый $\|\mathbf{x}\|$ и удовлетворяющий следующим условиям:

$$\|\mathbf{x}\| > 0, \text{если } \mathbf{x} \neq 0$$
$$\|\mathbf{x}\| = 0 \Leftrightarrow x = 0$$
$$\|\alpha\mathbf{x}\| = |\alpha| \, \|\mathbf{x}\|$$
$$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$$

Наиболее употребительны следующие нормы:

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{(\mathbf{x}, \mathbf{x})} - \text{Евклидова, или } l^2 \text{ норма}$$
$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i| - \text{Манхэттенская, или } l^1 \text{ норма}$$
$$\|\mathbf{x}\|_\infty = \max_{1 \leq i \leq n} |x_i| - \text{иногда называют нормой Чебышёва}$$

Опр. 23.2. Нормы $\|\cdot\|_I$ и $\|\cdot\|_{II}$ называются эквивалентными, если $\forall \mathbf{x} \in \mathbb{R}^n$ с одними и теми же фиксированными положительными постоянными c_1 и c_2 справедливо

$$c_1 \|\cdot\|_{II} \leq \|\cdot\|_I \leq c_2 \|\cdot\|_{II}$$

Пример 23.1. Найдём константы эквивалентности, связывающие нормы $\|\mathbf{x}\|_\infty, \|\mathbf{x}\|_1, \|\mathbf{x}\|_2$, а также векторы, на которых они достигаются (для которых выполняется равенство)

Доказательство.

$$\max_{1 \leq i \leq n} |x_i| \leq \sum_{i=1}^{n} |x_i| \leq n \max_{1 \leq i \leq n} |x_i| \Rightarrow \|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_1 \leq n \|\mathbf{x}\|_\infty \tag{1}$$

$$\begin{array}{l} \sum\limits_{i=1}^{n} x_i^2 \leq \left(\sum\limits_{i=1}^{n} |x_i| \right)^2 \\ \frac{\sum\limits_{i=1}^{n} |x_i|}{n} \leq \sqrt{\frac{\sum\limits_{i=1}^{n} x_i^2}{n}} \end{array} \Rightarrow \frac{1}{\sqrt{n}} \|\mathbf{x}\|_1 \leq \|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \tag{2}$$

$$\max_{1 \leq i \leq n} x_i^2 \leq \sum_{i=1}^{n} x_i^2 \leq n \max_{1 \leq i \leq n} x_i^2 \Rightarrow \|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \sqrt{n} \|\mathbf{x}\|_\infty \tag{3}$$

В полученных неравенствах константы эквивалентности достигаются на векторах либо с равными компонентами, либо с единственной ненулевой компонентой. □

Теорема 23.1. *Пусть B - симметричная положительно определенная матрица. Тогда можно принять за норму вектора $\mathbf{x}$ следующую величину*

$$\sqrt{(B\mathbf{x}, \mathbf{x})} = \|\mathbf{x}\|_B$$

и будет верна оценка

$$\sqrt{\min_i \lambda_i} \|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_B \leq \sqrt{\max_i \lambda_i} \|\mathbf{x}\|_2$$

где $\{\lambda_i\}$ - собственные числа матрицы B.

Доказательство. Проверим свойства нормы:

1. Положительность нормы для ненулевых векторов и равенство нулю для нулевого следует из определения,
$$\text{матрица } B \text{ положительно определена} \Leftrightarrow \forall \mathbf{x} \neq 0 \;\; \mathbf{x}^\intercal B \mathbf{x} = (B\mathbf{x}, \mathbf{x}) > 0$$
2. Вынесем скаляр α : $\sqrt{(B\alpha\mathbf{x}, \alpha\mathbf{x})} = \sqrt{\alpha^2 (B\mathbf{x}, \mathbf{x})} = |\alpha| \sqrt{(B\mathbf{x}, \mathbf{x})}$

3. Самым нетривиальным является проверка выполнимости неравенства треугольника. Для положительно определенной матрицы B $\exists C : B = C^2$. Поскольку она ещё и симметрична, то $B = B^\intercal \Rightarrow C = C^\intercal$

$$(B(\mathbf{x}+\mathbf{y}),\mathbf{x}+\mathbf{y}) = (\mathbf{x}+\mathbf{y})^\intercal B(\mathbf{x}+\mathbf{y}) = (\mathbf{x}+\mathbf{y})^\intercal C^2(\mathbf{x}+\mathbf{y}) = (\mathbf{x}+\mathbf{y})^\intercal C^\intercal C(\mathbf{x}+\mathbf{y}) =$$
$$= (C(\mathbf{x}+\mathbf{y}))^\intercal (C(\mathbf{x}+\mathbf{y})) = (C(\mathbf{x}+\mathbf{y}), C(\mathbf{x}+\mathbf{y})) = (C\mathbf{x}+C\mathbf{y}, C\mathbf{x}+C\mathbf{y}) = [C\mathbf{x}=\tilde{\mathbf{x}}, C\mathbf{y}=\tilde{\mathbf{y}}] =$$
$$= (\tilde{\mathbf{x}},\tilde{\mathbf{x}}) + (\tilde{\mathbf{y}},\tilde{\mathbf{y}}) + (\tilde{\mathbf{x}},\tilde{\mathbf{y}}) + (\tilde{\mathbf{y}},\tilde{\mathbf{x}}) \le (\tilde{\mathbf{x}},\tilde{\mathbf{x}}) + (\tilde{\mathbf{y}},\tilde{\mathbf{y}}) + |(\tilde{\mathbf{x}},\tilde{\mathbf{y}})| + |(\tilde{\mathbf{y}},\tilde{\mathbf{x}})| \le \|\tilde{\mathbf{x}}\|^2 + \|\tilde{\mathbf{y}}\|^2 + 2\|\tilde{\mathbf{x}}\|\,\|\tilde{\mathbf{y}}\| =$$
$$= (\|\tilde{\mathbf{x}}\| + \|\tilde{\mathbf{y}}\|)^2 = (\|C\mathbf{x}\| + \|C\mathbf{y}\|)^2 = (\sqrt{(C\mathbf{x},C\mathbf{x})} + \sqrt{(C\mathbf{y},C\mathbf{y})})^2 = (\sqrt{(B\mathbf{x},\mathbf{x})} + \sqrt{(B\mathbf{y},\mathbf{y})})^2$$

Итого имеем $\|\mathbf{x}+\mathbf{y}\|_B \le \|\mathbf{x}\|_B + \|\mathbf{y}\|_B$

Теперь докажем вторую часть. Так как $B = B^\intercal > 0$, то собственные векторы матрицы различны и ортогональны. Пусть $\mathbf{e}_1,\dots,\mathbf{e}_n-$ ортонормированная система собственных векторов матрицы B, (т.е. $(\mathbf{e}_i,\mathbf{e}_j) = \delta_i^j$), а $\lambda_1,\dots,\lambda_n-$ соответствующие собственные значения. Любой вектор $\mathbf{x}$ представим в виде $\mathbf{x} = \sum\limits_{i=1}^{n} c_i\mathbf{e}_i$. Поэтому

$$(B\mathbf{x},\mathbf{x}) = \left(\sum_{i=1}^{n}\lambda_i c_i\mathbf{e}_i, \sum_{i=1}^{n} c_i\mathbf{e}_i\right) = \sum_{i=1}^{n}\lambda_i c_i^2$$

Отсюда для произвольного вектора $\mathbf{x}$ имеем

$$\min_i \lambda_i(\mathbf{x},\mathbf{x}) \le (B\mathbf{x},\mathbf{x}) \le \max_i \lambda_i(\mathbf{x},\mathbf{x}), (\mathbf{x},\ \ \mathbf{x}) = \sum_{i=1}^{n} c_i^2$$

Так как все $\lambda_i(B) > 0$, то полученное неравенство означает эквивалентность евклидовой норме $\|\mathbf{x}\|_2$ с постоянными

$$\tilde{c_1} = \sqrt{\min_i \lambda_i},\ \ \tilde{c_2} = \sqrt{\max_i \lambda_i}$$

□

Опр. 23.3. Нормой матрицы A называется функционал, обозначаемый $\|A\|$ и удовлетворяющий следующим условиям:

$$\|A\| > 0, \text{если} A \ne 0$$
$$\|A\| = 0 \Leftrightarrow A = 0$$
$$\|\alpha A\| = |\alpha|\,\|A\|$$
$$\|A+B\| \le \|A\| + \|B\|$$
$$\|AB\| \le \|A\|\,\|B\|$$

Пример 23.2. Функционал $\|A\|_F = \sqrt{\sum\limits_{i,j=1}^{n} a_{ij}^2}$ является матричной нормой и называется нормой Фробениуса.

Пример 23.3. Функционал $\eta(A) = \max\limits_{i,j}|a_{ij}|$ (максимальный по модулю элемент матрицы) не является матричной нормой. Нарушается последнее свойство: рассмотрим матрицы $A = B : a_{ij} = b_{ij} = 1$. Тогда $\eta(A) = \eta(B) = 1$, но $\eta(AB) = \sum\limits_{i=1}^{n} 1 = n \Rightarrow \eta(AB) \not\le \eta(A)\eta(B)$. Тем не менее функционал $M(A) = n\eta(A)$ является матричной нормой.

$$M(AB) = n\max_{i,j}\left|\sum_{k=1}^{n} a_{ik}b_{kj}\right| \le n\max_{i,j}\sum_{k=1}^{n}|a_{ik}b_{kj}| \le n\sum_{k=1}^{n}\eta(A)\eta(B) = n\eta(A)n\eta(B) = M(A)M(B)$$

Лемма 23.1. *Пусть задана некоторая векторная норма $\|\cdot\|_v$. Тогда матричную норму можно определить как операторную*

$$\|A\|_v = \sup_{\|\mathbf{x}\|_v \ne 0} \frac{\|A\mathbf{x}\|_v}{\|\mathbf{x}\|_v} = \sup_{\|\mathbf{x}\|_v = 1} \|A\mathbf{x}\|_v$$

Доказательство. Доказательство сводится к проверке свойств матричной нормы. Для доказательства данного факта используется соотношение $\|A\mathbf{x}\| \le \|A\|\,\|\mathbf{x}\|$. Докажем его от противного. Пусть $\|\mathbf{x}\| \ne 0$ (иначе в следующем неравенстве сразу противоречие), предположим, что $\|A\mathbf{x}\| > \|A\|\,\|\mathbf{x}\|$, тогда из этого

следует $\frac{1}{\|\mathbf{x}\|}\|A\mathbf{x}\| > \|A\|$. Заметим, что $\frac{1}{\|\mathbf{x}\|}$ является скаляром, внесём его под норму (вектора). $\left\|A\frac{\mathbf{x}}{\|\mathbf{x}\|}\right\| > \|A\| = \sup\limits_{\|\mathbf{y}\|=1}\|A\mathbf{y}\|$. Но $\frac{\mathbf{x}}{\|\mathbf{x}\|}$ это вектор единичной длины, а значит имеем противоречие, т.к. получили что-то большее чем точная верхняя грань (справа по определению)

Теперь можно доказать последнее свойство матричной нормы. В соответствии с доказанным свойством имеем

$$\|AB\| = \sup_{\|\mathbf{x}\|\neq 0}\frac{\|AB\mathbf{x}\|}{\|\mathbf{x}\|} = \sup_{\|\mathbf{x}\|\neq 0}\frac{\|A(B\mathbf{x})\|}{\|\mathbf{x}\|} \leq \sup_{\|\mathbf{x}\|\neq 0}\frac{\|A\|\,\|B\mathbf{x}\|}{\|\mathbf{x}\|} \leq \sup_{\|\mathbf{x}\|\neq 0}\frac{\|A\|\,\|B\|\,\|\mathbf{x}\|}{\|\mathbf{x}\|} = \|A\|\,\|B\|$$

□

Опр. 23.4. Построенная матричная норма называется подчиненной соответствующей векторной норме $\|\cdot\|_v$.

Замечание. Для единичной матрицы I и произвольной подчиненной матричной нормы $\|I\|_v = 1$.

Теорема 23.2. *Матричные нормы, подчиненные векторным нормам $\|\cdot\|_\infty, \|\cdot\|_1, \|\cdot\|_2$ имеют вид*

$\|A\|_\infty = \max\limits_i \sum\limits_{j=1}^n |a_{ij}|$ *– максимум сумм строкам*

$\|A\|_1 = \max\limits_j \sum\limits_{i=1}^n |a_{ij}|$ *– максимум сумм по столбцам*

$\|A\|_2 = \sqrt{\max\limits_i \lambda_i(A^\intercal A)}$ *– корень из максимального собственного числа матрицы $A^\intercal A$*

Доказательство. Получим оценку сверху для $\|A\mathbf{x}\|_\infty$

$$\|A\mathbf{x}\|_\infty = \max_i\left|\sum_{j=1}^n a_{ij}x_j\right| \leq \max_i\left(\sum_{j=1}^n |a_{ij}|\max_j|x_j|\right) \leq \max_i\left(\sum_{j=1}^n |a_{ij}|\right)\|\mathbf{x}\|_\infty$$

Пусть максимальная построчная сумма соответствует строке с номером l. Тогда возьмём вектор $\mathbf{x}$, состоящий из знаков элементов соответствующей строки. Векторная норма будет равна 1, поэтому последнее неравенство обратиться в равенство. В результате умножения строки l на такой вектор $\mathbf{x}$ получим, фактически, сумму модулей, тем самым оставшееся неравенство также обратиться в равенство. Значит мы показали, что оценка достигается. Поделим на $\|\mathbf{x}\|_\infty$ и получим

$$\|A\|_\infty = \sup_{\|\mathbf{x}\|_\infty=1}\|A\mathbf{x}\|_\infty = \max_i\sum_{j=1}^n |a_{ij}|$$

Для вывода $\|A\mathbf{x}\|_1$ обозначим $A_1, \ldots, A_n$– столбцы матрицы. Заметим, что умножая матрицу слева на вектор-столбец справа, элемент столбца матрицы будет умножаться на одну и ту же компоненту вектора:

$$\|A\mathbf{x}\|_1 = \left\|\sum_{j=1}^n x_jA_j\right\|_1 \leq \sum_{j=1}^n\|x_jA_j\|_1 = \sum_{j=1}^n|x_j|\,\|A_j\|_1 \leq \max_j\|A_j\|_1\sum_{j=1}^n|x_j| = \max_j\|A_j\|_1\,\|\mathbf{x}\|_1$$

Получили желаемую оценку. Она достигается на векторе $\mathbf{e}_l$, где l– индекс максимального по норме столбца

Теперь для евклидовой нормы.

$$\|A\|_2 = \sup_{\mathbf{x}\neq 0}\frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \sup_{\mathbf{x}\neq 0}\sqrt{\frac{(A\mathbf{x},A\mathbf{x})}{(\mathbf{x},\mathbf{x})}}$$

Заметим, что $(A^\intercal A)^\intercal = A^\intercal(A^\intercal)^\intercal = (A^\intercal A)$, т.е матрица $B = A^\intercal A$– симметричная. Также имеем $0 \leq (A\mathbf{x},A\mathbf{x}) = (A\mathbf{x})^\intercal(A\mathbf{x}) = \mathbf{x}^\intercal A^\intercal A\mathbf{x} = (A^\intercal A\mathbf{x},\mathbf{x}) = (B\mathbf{x},\mathbf{x})$, следовательно, все $\lambda(B) \geq 0$. Тогда

$$\sup_{\mathbf{x}\neq 0}\sqrt{\frac{(A\mathbf{x},A\mathbf{x})}{(\mathbf{x},\mathbf{x})}} = \sup_{\mathbf{x}\neq 0}\sqrt{\frac{(A^\intercal A\mathbf{x},\mathbf{x})}{(\mathbf{x},\mathbf{x})}} = \sup_{\mathbf{x}\neq 0}\sqrt{\frac{(B\mathbf{x},\mathbf{x})}{(\mathbf{x},\mathbf{x})}} \leq \sqrt{\max_i\lambda_i(B)}$$

(Последнее следует из эквивалентности данной нормы евклидовой, а именно из $(B\mathbf{x},\mathbf{x}) \leq \max\limits_i\lambda_i\|\mathbf{x}\|_2$)
Равенство достигается на соответствующем собственном векторе. Поэтому $\|A\|_2 = \sqrt{\max\limits_i\lambda_i(A^\intercal A)}$

Важный частный случай симметричной матрицы $A = A^\intercal$. Тогда $\|A\|_2 = \max\limits_i|\lambda_i(A)|$ □

Теорема 23.3. *Модуль любого собственного значения матрицы не больше любой ее нормы:* $|\lambda(A)| \le \|A\|$

Доказательство. Зафиксируем произвольный собственный вектор $\mathbf{x}$ матрицы A и построим матрицу X, столбцами которой являются вектора $\mathbf{x}$. Получим равенство $\lambda X = AX$. Отсюда следует

$$|\lambda| \, \|X\| = \|\lambda X\| = \|AX\| \le \|A\| \, \|X\|$$

□

Следствие 23.1. Для любого собственного значения $\lambda(A)$ невырожденной матрицы A справедлива оценка $\frac{1}{\|A^{-1}\|} \le |\lambda(A)|$

Опр. 23.5. Величина $\mathrm{cond}(A) = \|A\| \, \|A^{-1}\|$ называется числом обусловленности матрицы A. Для вырожденных матриц $\mathrm{cond}(A) = \infty$. Конкретное значение $\mathrm{cond}(A)$ зависит от выбора матричной нормы, однако, в силу их эквивалентности при практических оценках этим различием обычно можно пренебречь. Если $\mathrm{cond}(A)$ велико, то матрицу называют *плохо обусловленной*.

Рассмотрим систему линейных алгебраических уравнений $A\mathbf{x} = \mathbf{b}$ с квадратной невырожденной матрицей A и точным решением $\mathbf{x}$. В результате численного решения с конечной разрядностью вместо $\mathbf{x}$ получается *приближенное* решение $\tilde{\mathbf{x}} : A\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$. При этом вектор $\mathbf{z} = \mathbf{x} - \tilde{\mathbf{x}}$ называют *вектор ошибки*, а вектор $\mathbf{r} = \mathbf{b} - A\tilde{\mathbf{x}}$ - *вектор невязки*.

Найдем насколько приближенное решение отличается от точного. Из неравенств

$$\|\mathbf{z}\| = \|\mathbf{x} - \tilde{\mathbf{x}}\| \le \|A^{-1}\| \left\|\mathbf{b} - \tilde{\mathbf{b}}\right\|, \|A\| \, \|\mathbf{x}\| \ge \|\mathbf{b}\|$$

получаем, что для относительной ошибки верна оценка (во втором неравенстве нужно единицу поделить на левую и правую части, в результате чего неравенство поменяет знак. После домножения на норму матрицы A нужно это всё перемножить с первым неравенством)

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \le \|A\| \, \|A^{-1}\| \frac{\left\|\mathbf{b} - \tilde{\mathbf{b}}\right\|}{\|\mathbf{b}\|} = \|A\| \, \|A^{-1}\| \frac{\|\mathbf{b} - A\tilde{\mathbf{x}}\|}{\|\mathbf{b}\|} = \mathrm{cond}(A) \frac{\|\mathbf{b} - A\tilde{\mathbf{x}}\|}{\|\mathbf{b}\|}$$

Пример 23.4. Невязка и ошибка не одно и то же. Проиллюстрируем это на примере матрицы с плохой обусловленностью. Тогда невязка не может гарантировать малость относительной ошибки. Более того, может оказаться так, что достаточно точное решение будет иметь большую невязку. Пусть

$$\begin{pmatrix} \varepsilon & 0 \\ 0 & \varepsilon^{-1} \end{pmatrix} \begin{pmatrix} 1 \\ \varepsilon \end{pmatrix} = \begin{pmatrix} \varepsilon \\ 1 \end{pmatrix}, \ \ \varepsilon \ll 1$$

Вектор $\tilde{\mathbf{x}} = (1 + 1, \varepsilon)^\intercal$, который не является близким к $\mathbf{x}$, дает маленькую невязку $\mathbf{r} = (-\varepsilon, 0)^\intercal$. Вектор $\tilde{\mathbf{x}} = (1, \varepsilon + \sqrt{\varepsilon})^\intercal$ достаточно близок к $\mathbf{x}$ в смысле относительной погрешности, однако $\tilde{\mathbf{x}}$ дает большую невязку $\mathbf{r} = (0, \frac{-1}{\sqrt{\varepsilon}})^\intercal$.

Лемма 23.2.

$$\forall A \ \ \mathrm{cond}(\alpha A) = \mathrm{cond}(A) \tag{4}$$

$$\mathrm{cond}(A) \ge 1 \tag{5}$$

$$\forall A = A^T \ \ \mathrm{cond}_2(A) = \frac{\max_i \lambda_i(A)}{\min_i \lambda_i(A)} \tag{6}$$

$$\forall Q : \ Q^T = Q^{-1} \Rightarrow \mathrm{cond}_2(Q) = 1 \tag{7}$$

Доказательство.

(4) $\mathrm{cond}(\alpha A) = \|\alpha A\| \left\|\alpha^{-1} A^{-1}\right\| = \frac{|\alpha|}{|\alpha|} \mathrm{cond}(A)$

(5) $\mathrm{cond}(A) = \|A\| \left\|A^{-1}\right\| \ge \left\|AA^{-1}\right\| = \|I\| \ge 1$

(6) $\mathrm{cond}_2(A) = \|A\| \left\|A^{-1}\right\| = \max\limits_i \lambda_i(A) \cdot \max\limits_i \lambda_i^{-1}(A) = \frac{\max_i \lambda_i(A)}{\min_i \lambda_i(A)}$

(7) $\|Q\|_2 = \sup\limits_{\|\mathbf{x}\|_2 \ne 0} \frac{\sqrt{(Q\mathbf{x}, Q\mathbf{x})}}{\|\mathbf{x}\|_2} = \sup\limits_{\|\mathbf{x}\|_2 \ne 0} \frac{\sqrt{(Q^\intercal Q\mathbf{x}, \mathbf{x})}}{\|\mathbf{x}\|_2} = 1 =$

$$\sup_{\|\mathbf{x}\|_2 \ne 0} \frac{\sqrt{(QQ^\intercal\mathbf{x}, \mathbf{x})}}{\|\mathbf{x}\|_2} = \sup_{\|\mathbf{x}\|_2 \ne 0} \frac{\sqrt{(Q^\intercal\mathbf{x}, Q^\intercal\mathbf{x})}}{\|\mathbf{x}\|_2} = \|Q^\intercal\|_2 = \left\|Q^{-1}\right\|_2$$

□

Так же обусловленность не связана с определителем матрицы, приведем примеры

Пример 23.5. Покажем, что если определитель матрицы мал, то матрица не обязательно плохо обусловлена, а определитель плохо обусловленной матрицы может равняться 1.

Пусть дана диагональная матрица $A = \varepsilon I$, где $\varepsilon > 0-$ малое число и $I-$ единичная матрица. Определитель $\det(A) = \varepsilon^n$ весьма мал, тогда как матрица A хорошо обусловлена, поскольку

$$\operatorname{cond}_\infty(A) = \|A\|_\infty \left\|A^{-1}\right\|_\infty = \varepsilon \|I\|_\infty \varepsilon^{-1} \left\|A^{-1}\right\|_\infty = 1$$

$$\text{Пусть } A = \begin{pmatrix} \varepsilon & 0 \\ 0 & \varepsilon^{-1} \end{pmatrix}, \text{ тогда } \det(A) = 1, \ \operatorname{cond}_\infty(A) = \varepsilon^{-2}$$

Пример 23.6. Пусть

$$A = \begin{pmatrix} 1 & -1 & -1 & \dots & -1 \\ 0 & 1 & -1 & \dots & -1 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}, \ A^{-1} = \begin{pmatrix} 1 & 1 & 2 & \dots & 2^{n-2} \\ 0 & 1 & 1 & \dots & 2^{n-3} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$

$\det(A) = 1$, $\|A\|_\infty = n$, $\left\|A^{-1}\right\|_\infty = 1 + 1 + 2 + 2^2 + \ldots + 2^{n-2} = 2^{n-1}$, $\operatorname{cond}_\infty(A) = n2^{n-1}$. Т.е. матрица плохо обусловлена, хотя $\det(A) = 1$. Отметим, что $\lambda_i(A) = 1$, но матрица A несимметрична, поэтому за обусловленность отвечают $\lambda(A^\intercal A)$.

24 Метод Гаусса решения систем линейных алгебраических уравнений. Алгоритм ортогонализации Грама–Шмидта.

Метод Гаусса решения систем линейных алгебраических уравнений

К точным методам решения системы $A\mathbf{x} = \mathbf{b}$ линейных алгебраических уравнений (СЛАУ) относятся алгоритмы, которые при отсутствии ошибок округления позволяют точно вычислить искомый вектор $\mathbf{x}$. Если число ненулевых элементов матрицы имеет порядок n^2, то большинство такого рода алгоритмов позволяет найти решение за $\underline{\underline{\mathcal{O}}}(n^3)$ арифметических действий. Данная оценка, а также необходимость хранения всех элементов матрицы в памяти машины, накладывают существенное ограничение на область применимости точных методов. Однако, для решения задач не очень большой размерности (от 10^3 до 10^4) в большинстве случаев разумно применение точных алгоритмов. Отметим, что при численном решении задач математической физики часто требуется обращать матрицы блочнодиагонального вида. В этом случае удается построить точные методы с меньшим по порядку числом арифметических действий. К таким алгоритмам относятся метод прогонки, стрельбы, Фурье (базисных функций).

Метод исключения Гаусса является наиболее известным из точных методов, применяемых для задач с матрицами общего вида. В предположении, что коэффициент $a_{11} \neq 0$, в первом шаге уравнения исходной системы заменяются на следующие

$$\begin{cases} x_1 + \sum\limits_{j=2}^{n} \dfrac{a_{1j}}{a_{11}} x_j = \dfrac{b_1}{a_{11}} \\ \sum\limits_{j=2}^{n} x_j \left(a_{ij} - \dfrac{a_{i1}}{a_{11}} a_{1j} \right) = b_i - \dfrac{b_1}{a_{11}} a_{i1}, \ i = 2, \ldots, n \end{cases}$$

В матричном виде при $n = 3$ такой переход выглядит следующим образом. Обозначим $A^{(0)} \equiv A$, $\mathbf{b}^{(0)} \equiv \mathbf{b}$

$$A^{(0)}\mathbf{x} = \mathbf{b}^{(0)} \Leftrightarrow \left(\begin{array}{ccc|c} 1 & \frac{a_{12}}{a_{11}} & \frac{a_{13}}{a_{11}} & \frac{b_1}{a_{11}} \\ 0 & a_{22} - \frac{a_{12}}{a_{11}} a_{21} & a_{23} - \frac{a_{12}}{a_{11}} a_{21} & b_2 - \frac{b_1}{a_{11}} a_{21} \\ 0 & a_{32} - \frac{a_{12}}{a_{11}} a_{31} & a_{33} - \frac{a_{12}}{a_{11}} a_{31} & b_3 - \frac{b_1}{a_{11}} a_{31} \end{array} \right) =: \left(\begin{array}{ccc|c} 1 & \overline{a_{12}} & \overline{a_{13}} & \overline{b_1} \\ 0 & \overline{a_{22}} & \overline{a_{23}} & \overline{b_2} \\ 0 & \overline{a_{32}} & \overline{a_{33}} & \overline{b_3} \end{array} \right) \Leftrightarrow A^{(1)}\mathbf{x} = \mathbf{b}^{(1)}$$

то есть первое уравнение делится на a_{11}, а затем, умноженное на соответствующий коэффициент a_{i1}, вычитается из последующих уравнений. В полученной системе $A^{(1)}\mathbf{x} = \mathbf{b}^{(1)}$ неизвестное x_1 оказывается исключенным из всех уравнений, кроме первого. Далее, при условии, что коэффициент $a_{22}^{(1)}$ матрицы $A^{(1)}$ отличен от нуля, исключаем x_2 из всех уравнений кроме первого и второго, и т.д.

$$A^{(1)}\mathbf{x} = \mathbf{b}^{(1)} \to \left(\begin{array}{ccc|c} 1 & \overline{a_{12}} & \overline{a_{13}} & \overline{b_1} \\ 0 & 1 & \frac{\overline{a_{23}}}{\overline{a_{22}}} & \frac{\overline{b_2}}{\overline{a_{22}}} \\ 0 & 0 & \overline{a_{33}} - \frac{\overline{a_{23}}}{\overline{a_{22}}} \overline{a_{32}} & \overline{b_3} - \frac{\overline{b_2}}{\overline{a_{22}}} \overline{a_{32}} \end{array} \right) =: \left(\begin{array}{ccc|c} 1 & \widehat{a_{12}} & \widehat{a_{13}} & \widehat{b_1} \\ 0 & 1 & \widehat{a_{23}} & \widehat{b_2} \\ 0 & 0 & \widehat{a_{33}} & \widehat{b_3} \end{array} \right) \Leftrightarrow A^{(2)}\mathbf{x} = \mathbf{b}^{(2)}$$

В итоге получим систему $A^{(n)}\mathbf{x} = \mathbf{b}^{(n)}$ с верхнетреугольной матрицей.

$$A^{(2)}\mathbf{x} = \mathbf{b}^{(2)} \to \left(\begin{array}{ccc|c} 1 & \widehat{a_{12}} & \widehat{a_{13}} & \widehat{b_1} \\ 0 & 1 & \widehat{a_{23}} & \widehat{b_2} \\ 0 & 0 & 1 & \frac{\widehat{b_3}}{\widehat{a_{33}}} \end{array} \right) \Leftrightarrow A^{(3)}\mathbf{x} = \mathbf{b}^{(3)}$$

Данная последовательность вычислений называется *прямым ходом метода Гаусса*. Из последнего уравнения приведенной системы определяем компоненту решения $\mathbf{x}$. Далее подставляем x_n в $(n-1)$-е уравнение, находим x_{n-1} и т.д. Соответствующая последовательность вычислений называется *обратным ходом Гаусса*.

$$\left(\begin{array}{ccc|c} 1 & \widehat{a_{12}} & \widehat{a_{13}} & \widehat{b_1} \\ 0 & 1 & \widehat{a_{23}} & \widehat{b_2} \\ 0 & 0 & 1 & \frac{\widehat{b_3}}{\widehat{a_{33}}} \end{array} \right) \to \left(\begin{array}{ccc|c} 1 & 0 & 0 & \widehat{b_1} - \frac{\widehat{b_3}}{\widehat{a_{33}}} \widehat{a_{13}} - \left(\widehat{b_2} - \frac{\widehat{b_3}}{\widehat{a_{33}}} \widehat{a_{23}} \right) \widehat{a_{12}} \\ 0 & 1 & 0 & \widehat{b_2} - \frac{\widehat{b_3}}{\widehat{a_{33}}} \widehat{a_{23}} \\ 0 & 0 & 1 & \frac{\widehat{b_3}}{\widehat{a_{33}}} \end{array} \right)$$

Если на k-м шаге прямогохода коэффициента $a_{kk}^{(k-1)}$ равен нулю, тогда k-я строка уравнения переставляется с произвольной l-й строкой, $l > k$ с ненулевым коэффициентом $a_{lk}^{(k-1)}$ при x_k. Такая строка всегда найдется, если $\det(A) \neq 0$.

Если на k-м шаге прямого хода диагональный элемент $a_{kk}^{(k-1)}$ отличен от нуля, но имеет малое абсолютное значение, то коэффициенты очередной матрицы $A^{(k)}$ будут вычислены с большой абсолютной

погрешностью. Это может существенно исказить найденный ответ. Поэтому при практической реализации метода Гаусса следует на каждом шаге прямого хода переставлять на k-е место строку с максимальным по модулю элементом $a_{lk}^{(k-1)}$ среди всех $l \geq k$. Такая необходимая при расчетах модификация назвется методом Гаусса с частичным выбором главного элемента. Данный алгоритм позволяет гарантированно найти приближенное решение $\mathbf{x}$ с малой нормой невязки $\dfrac{\|\mathbf{b} - A\tilde{\mathbf{x}}\|}{\|\tilde{\mathbf{x}}\|}$, но, возможно, с большой нормой ошибки $\|\mathbf{x} - \tilde{\mathbf{x}}\|$.

Пример 24.1. Пример нахождения решения СЛАУ методом Гаусса с выбором главного элемента по столбцу.

$$\left(\begin{array}{ccc|c} 1.00 & 2.00 & 3.00 & 3.00 \\ 3.00 & 1.00 & 2.00 & 6.00 \\ 2.00 & 3.00 & 1.00 & 9.00 \end{array}\right) \to \left(\begin{array}{ccc|c} 3.00 & 1.00 & 2.00 & 6.00 \\ 1.00 & 2.00 & 3.00 & 3.00 \\ 2.00 & 3.00 & 1.00 & 9.00 \end{array}\right) \to \left(\begin{array}{ccc|c} 1.00 & 0.33 & 0.67 & 2.00 \\ 1.00 & 2.00 & 3.00 & 3.00 \\ 2.00 & 3.00 & 1.00 & 9.00 \end{array}\right) \to$$

$$\left(\begin{array}{ccc|c} 1.00 & 0.33 & 0.67 & 2.00 \\ 0.00 & 1.67 & 2.33 & 1.00 \\ 0.00 & 2.33 & -0.33 & 5.00 \end{array}\right) \to \left(\begin{array}{ccc|c} 1.00 & 0.33 & 0.67 & 2.00 \\ 0.00 & 2.33 & -0.33 & 5.00 \\ 0.00 & 1.67 & 2.33 & 1.00 \end{array}\right) \to \left(\begin{array}{ccc|c} 1.00 & 0.33 & 0.67 & 2.00 \\ 0.00 & 1.00 & -0.14 & 2.14 \\ 0.00 & 1.67 & 2.33 & 1.00 \end{array}\right) \to$$

$$\left(\begin{array}{ccc|c} 1.00 & 0.33 & 0.67 & 2.00 \\ 0.00 & 1.00 & -0.14 & 2.14 \\ 0.00 & 0.00 & 2.57 & -2.57 \end{array}\right) \to \left(\begin{array}{ccc|c} 1.00 & 0.33 & 0.67 & 2.00 \\ 0.00 & 1.00 & -0.14 & 2.14 \\ 0.00 & 0.00 & 2.57 & -2.57 \end{array}\right) \to \left(\begin{array}{ccc|c} 1.00 & 0.33 & 0.67 & 2.00 \\ 0.00 & 1.00 & -0.14 & 2.14 \\ 0.00 & 0.00 & 1.00 & -1.00 \end{array}\right) \to$$

$$\left(\begin{array}{ccc|c} 1.00 & 0.33 & 0.00 & 2.67 \\ 0.00 & 1.00 & 0.00 & 2.00 \\ 0.00 & 0.00 & 1.00 & -1.00 \end{array}\right) \to \left(\begin{array}{ccc|c} 1.00 & 0.00 & 0.00 & 2.00 \\ 0.00 & 1.00 & 0.00 & 2.00 \\ 0.00 & 0.00 & 1.00 & -1.00 \end{array}\right) \to \left(\begin{array}{ccc|c} 1.00 & 0.00 & 0.00 & 2.00 \\ 0.00 & 1.00 & 0.00 & 2.00 \\ 0.00 & 0.00 & 1.00 & -1.00 \end{array}\right)$$

Лемма 24.1. *Реализация прямого и обратного хода метода Гаусса требует порядка $\frac{2}{3}n^3$ и n^2 арифметических действий соответственно.*

Доказательство.

$$\left(\begin{array}{c|c} 1 & \\ \hline 0 & \\ \vdots & (n-1) \times (n-1) \\ 0 & \end{array}\right)$$

$n \times n$

За первый шаг прямого хода метода Гаусса мы делаем n^2 операций сложения и умножения: первый столбец зануляем, остальные умножаем на соответствующие коэффициенты. За каждый следующий шаг метода мы будем работать с меньшим квадратом $(n-1) \times (n-1)$.

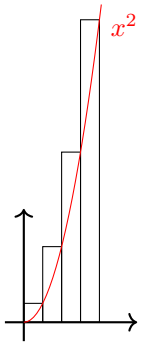
Таким образом, суммарно количество одной из операций за прямой ход метода Гаусса достигнет

$$n^2 + (n-1)^2 + \ldots + 1^2 = \sum_{k=1}^{n} k^2$$

Посчитаем это значение двумя способами:

$$\begin{aligned} (k+1)^3 &= k^3 + 3k^2 + 3k + 1 \\ 2^3 &= 1^3 + 3 \cdot 1^2 + 3 \cdot 1 + 1 \\ 3^3 &= 2^3 + 3 \cdot 2^2 + 3 \cdot 2 + 1 \Rightarrow \\ &\ldots \\ (n+1)^3 &= n^3 + 3 \cdot n^2 + 3 \cdot n + 1 \end{aligned}$$

$$\begin{aligned} \sum_{k=2}^{n+1} k^3 &= \sum_{k=1}^{n} k^3 + 3\left(\sum_{k=1}^{n} k^2\right) + 3\left(\sum_{k=1}^{n} k\right) + n \\ (n+1)^3 &= 1 + 3\left(\sum_{k=1}^{n} k^2\right) + 3\frac{(n+1)}{2}n + n \\ 3\left(\sum_{k=1}^{n} k^2\right) &= \frac{(n+1)(2n^2 + 4n - 3n)}{2} \\ \sum_{k=1}^{n} k^2 &= \frac{n(n+1)(2n+1)}{6} \approx \frac{n^3}{3} \end{aligned}$$



Получим это значение иначе. Представим эту сумму как площадь следующих фигур. Заметим, что искомое значение является формулой вычисления интеграла через прямоугольник по крайней точке.

$$\sum_{k=1}^{n} 1 \cdot k^2 \approx \int_0^n x^2 dx = \frac{n^3}{3}$$

За первый шаг обратного хода метода Гаусса мы делаем 1 операцию сложения и умножения: подставляем x_n, умножаем его на a_{n-1n} и вычитаем из b_{n-1}. За каждый следующий шаг метода мы будем подставлять все предыдущие x_i умножать на соответствующие a_{ij} и вычитать из соответствующего b_i. В итоге общее количество одной арифметической операции

$$1 + 2 + \ldots + (n-1) = (n-1)\frac{1+n-1}{2} = \frac{n^2-n}{2} \approx \frac{n^2}{2}$$

□

Теорема 24.1. *Пусть матрица $A_{kk}^{(k-1)}$ не требует перестановок $\forall\ k = 1, \ldots, n$. Тогда алгоритм Гаусса представим в виде*

$$C_n \ldots C_2' C_2 C_1' C_1 A x = C_n \ldots C_1' C_1 b$$

где C_i и C_i' матрицы вида

$$C_i = \begin{pmatrix} 1 & & & & & 0 \\ & \ddots & & & & \\ & & (a_{ii}^{(i-1)})^{-1} & & & \\ & & & \ddots & & \\ 0 & & & & & 1 \end{pmatrix}; \quad C_i' = \begin{pmatrix} 1 & & & & & & 0 \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & 1 & & & \\ & & & -a_{i+1,i}^{(i-1)} & 1 & & \\ & & & -a_{i+2,i}^{(i-1)} & 0 & \ddots & \\ & & & \vdots & & & \ddots \\ 0 & & & -a_{n,i}^{(i-1)} & & & 1 \end{pmatrix}$$

Доказательство. Доказывается непосредственной проверкой, что на i-ом шаге матрица $A^{(i-1)}$ при умножении слева на C_i нормирует первый элемент, а при умножении слева на C_i' из $i+1$ строки вычитается i-ая строка умноженная на $-a_{i+1,i}^{(i-1)}$. □

Метод Гаусса по сути соответствует разложению исходной матрицы A на произведение нижнетреугольной L и верхнетреугольной R. Действительно, $C = AR$, $C \stackrel{\text{def}}{=} C_n \ldots C_1' C_1$ - нижнетреугольная матрица, R - верхнетреугольная матрица с единичной диагональю. Поэтому $A = LR$, где $L = C^{-1}$.

Прямой ход метода Гаусса с частичным выбором главного элемента соответствует последовательному умножению исходной системы на некоторые диагональные матрицы C_k, нижнетреугольные матрицы C_k' и матрицы перестановок P_k^{ij}. При этом задание матриц C_k, C_k' совпадают с матриц метода Гаусса, матрицы P_k^{ij} получаются из единичной матрицы I некоторой перестановкой строк.

$$P_k^{ij} = \begin{pmatrix} 1 & & & & & & 0 \\ & \ddots & & & & & \\ & & 0 & & 1 & & \\ & & & \ddots & & & \\ & & 1 & & 0 & & \\ & & & & & \ddots & \\ 0 & & & & & & 1 \end{pmatrix} \begin{matrix} \\ \\ i \\ \\ j \\ \\ \\ \end{matrix}$$

Такой подход гарантирует, что метод находит приближенное решение с малой относительной невязкой (но, возможно, большой ошибкой). Метод Гаусса с выбором на каждом шаге наибольшего элемента по всей подматрице на практике почти не применяется, хотя имеются немногочисленные примеры, когда он дает качественное улучшение.

Отметим, что для невырожденной матрицы A существуют матрицы перестановок P_1 и P_2, нижняя треугольная матрица L и верхнетреугольная R такие, что $P_1 A P_2 = LR$. При этом достаточно оодной из матриц P_i. Умножение A на матрицу P_i слева переставляет строки исходной матрицы, а умножение на P_2 справа столбцы. Для того чтобы матрица имела LR-разложение, необходимо и достаточно, чтобы все ее ведущие подматрицы (в том числе и A) были невырожденные.

Алгоритм ортогонализации Грама-Шмидта

Хотим по набору линейно-независимых векторов $< \mathbf{a}_1, \ldots, \mathbf{a}_n >$ получить подпространство $< \mathbf{q}_1, \ldots, \mathbf{q}_n >$ такое, что $(\mathbf{q}_i, \mathbf{q}_j) = \delta_{ij}$.

Суть метода Грама-Шмидта состоит во взятии первого ортогонального вектора равным первому исходному вектору и построении каждого нового ортогонального вектора равным текущему исходному вектору, скорректированному на величины проекций текущего вектора на предыдущие ортогональные векторы.

$$\begin{array}{lcll} \widetilde{\mathbf{q}}_1 & = & \mathbf{a}_1 & \mathbf{q}_1 = \widetilde{\mathbf{q}}_1 / \left\|\widetilde{\mathbf{q}}_1\right\|_2 \\ \widetilde{\mathbf{q}}_2 & = & \mathbf{a}_2 - (\mathbf{q}_1, \mathbf{a}_2)\mathbf{q}_1 & \mathbf{q}_2 = \widetilde{\mathbf{q}}_2 / \left\|\widetilde{\mathbf{q}}_2\right\|_2 \\ \widetilde{\mathbf{q}}_3 & = & \mathbf{a}_3 - (\mathbf{q}_2, \mathbf{a}_3)\mathbf{q}_2 - (\mathbf{q}_1, \mathbf{a}_3)\mathbf{q}_1 & \mathbf{q}_3 = \widetilde{\mathbf{q}}_3 / \left\|\widetilde{\mathbf{q}}_3\right\|_2 \\ & \ldots & & \\ \widetilde{\mathbf{q}}_n & = & \mathbf{a}_n - \sum_{i=1}^{n-1}(\mathbf{q}_i, \mathbf{a}_n)\mathbf{q}_i & \mathbf{q}_n = \widetilde{\mathbf{q}}_n / \left\|\widetilde{\mathbf{q}}_n\right\|_2 \end{array}$$

Теорема 24.2. *Вектора* $\{\mathbf{q}_i\}_{i=1}^n$ *ортонормальны:* $(\mathbf{q}_i, \mathbf{q}_j) = \delta_{ij}$.

Доказательство. Проведем доказательство с помощью индукции

- База:

$$(\mathbf{q}_1, \mathbf{q}_1) = \left(\frac{\mathbf{a}_1}{\|\mathbf{a}_1\|_2}, \frac{\mathbf{a}_1}{\|\mathbf{a}_1\|_2}\right) = \frac{1}{\|\mathbf{a}_1\|_2^2}(\mathbf{a}_1, \mathbf{a}_1) = 1$$

$$(\mathbf{q}_1, \mathbf{q}_2) = \left(\frac{\mathbf{a}_1}{\|\mathbf{a}_1\|_2}, \frac{\mathbf{a}_2 - (\mathbf{q}_1, \mathbf{a}_2)\mathbf{q}_1}{\|\mathbf{a}_2 - (\mathbf{q}_1, \mathbf{a}_2)\mathbf{q}_1\|_2}\right) = \frac{(\mathbf{a}_1, (\mathbf{a}_1, \mathbf{a}_1)\mathbf{a}_2 - (\mathbf{a}_1, \mathbf{a}_2)\mathbf{a}_1)}{\|\mathbf{a}_1\|_2 \|(\mathbf{a}_1, \mathbf{a}_1)\mathbf{a}_2 - (\mathbf{a}_1, \mathbf{a}_2)\mathbf{a}_1\|_2} =$$

$$= \frac{(\mathbf{a}_1, (\mathbf{a}_1, \mathbf{a}_1)\mathbf{a}_2) - (\mathbf{a}_1, (\mathbf{a}_1, \mathbf{a}_2)\mathbf{a}_1)}{\|\mathbf{a}_1\|_2 \|(\mathbf{a}_1, \mathbf{a}_1)\mathbf{a}_2 - (\mathbf{a}_1, \mathbf{a}_2)\mathbf{a}_1\|_2} = \frac{(\mathbf{a}_1, \mathbf{a}_1)(\mathbf{a}_1, \mathbf{a}_2) - (\mathbf{a}_1, \mathbf{a}_2)(\mathbf{a}_1, \mathbf{a}_1)}{\|\mathbf{a}_1\|_2 \|(\mathbf{a}_1, \mathbf{a}_1)\mathbf{a}_2 - (\mathbf{a}_1, \mathbf{a}_2)\mathbf{a}_1\|_2} = 0$$

- Шаг индукции. Пусть предположение верно $\forall\ k < n$. Для упрощения выкладок рассмотрим ненормированный $\widetilde{\mathbf{q}_n}$.

$$(\widetilde{\mathbf{q}_n}, \mathbf{q}_k) = (\mathbf{a}_n - \sum_{i=1}^{n-1}(\mathbf{q}_i, \mathbf{a}_n)\mathbf{q}_i, \mathbf{q}_k) = (\mathbf{a}_n, \mathbf{q}_k) - \left(\sum_{i=1}^{n-1}(\mathbf{q}_i, \mathbf{a}_n)\mathbf{q}_i, \mathbf{q}_k\right) = (\mathbf{a}_n, \mathbf{q}_k) - (\mathbf{a}_n, \mathbf{q}_k)(\mathbf{q}_k, \mathbf{q}_k) = 0$$

□

Модифицированный алгоритм ортогонализации Грама-Шмидта

математически эквивалентен предыдущему, но более устойчив при наличии вычислительной погрешности:

$$\begin{array}{lcll} \mathbf{q}_1 & = & \mathbf{a}_1 & \mathbf{q}_1 = \mathbf{q}_1 / \left\|\mathbf{q}_1\right\|_2 \\ \mathbf{q}_2 & = & \mathbf{a}_2 & \\ \mathbf{q}_2 & = & \mathbf{q}_2 - (\mathbf{q}_1, \mathbf{q}_2)\mathbf{q}_1 & \mathbf{q}_2 = \mathbf{q}_2 / \left\|\mathbf{q}_2\right\|_2 \\ & \ldots & & \\ \mathbf{q}_n & = & \mathbf{a}_n & \\ \mathbf{q}_n & = & \mathbf{q}_n - (\mathbf{q}_1, \mathbf{q}_n)\mathbf{q}_1 & \\ \mathbf{q}_n & = & \mathbf{q}_n - (\mathbf{q}_2, \mathbf{q}_n)\mathbf{q}_2 & \\ & \ldots & & \\ \mathbf{q}_n & = & \mathbf{q}_n - (\mathbf{q}_{n-1}, \mathbf{q}_n)\mathbf{q}_{n-1} & \mathbf{q}_n = \mathbf{q}_n / \left\|\mathbf{q}_n\right\|_2 \end{array}$$

Вычислительная устойчивость достигается за счет вычитания уже наайденных $\mathbf{q}_i$ из $\mathbf{a}_n$. В этом случае накопление погрешности при вычислении $\mathbf{q}_n$ из-за неточной ортогональности $\{\mathbf{q}_i\}$, $i < n$ происходит существенно медленнее.

25 Метод отражений.

Среди точных методов, требующих для реализации порядка $\underline{\underline{\mathcal{O}}}(n^3)$ действий, одним из наиболее устойчивых к вычислительной погрешности является *метод отражений.*

Матрица Хаусхолдера

Пусть имеется некоторый единичный вектор $\mathbf{w} \in \mathbb{R}^n$, $\|\mathbf{w}\|_2 = 1$. Построим по нему следующую матрицу $U_{\mathbf{w}} = I - 2\mathbf{w}\mathbf{w}^T$, называемую матрицей Хаусхолдера. Здесь I - единичный оператор, $\mathbf{w}\mathbf{w}^T =: \Omega$ - матрица с элементами $w_{ij} = \mathbf{w}_i\mathbf{w}_j$, являющаяся результатом произведения вектор-столбца $\mathbf{w}$ на вектор-строку $\mathbf{w}^T$.

Теорема 25.1. *Матрица Хаусхолдера обладает следующими свойствами*

1. $U_{\mathbf{w}}$ *является симметричной и ортогональной, то есть* $U_{\mathbf{w}} = U_{\mathbf{w}}^T$, $U_{\mathbf{w}} = U_{\mathbf{w}}^{-1}$, $\lambda_i(U_{\mathbf{w}}) = \pm 1$
2. $U_{\mathbf{w}}\mathbf{w} = -\mathbf{w}$
3. $\mathbf{v} \perp \mathbf{w} \Rightarrow U_{\mathbf{w}}\mathbf{v} = \mathbf{v}$
4. *Образ оператора* $U_{\mathbf{w}}$ *на вектор из* $\mathbb{R}^n$ *является зеркальным отражением относительно гиперплоскости, перпендикулярной* $\mathbf{w}$.

Доказательство. 1. Симметричность матрицы $U_{\mathbf{w}}$ следует из явного вида $U_{\mathbf{w}}$. Так как $(\mathbf{w}, \mathbf{w})_2 = 1$, то

$$(\Omega\Omega)_{ij} = \left(\sum_{k=1}^{n}(\Omega)_{ik}(\Omega)_{kj}\right)_{ij} = \left(\sum_{k=1}^{n} w_i w_k w_k w_j\right)_{ij} = \left(w_i w_j \sum_{k=1}^{n} w_k^2\right)_{ij} = (w_i w_j)_{ij} = (\Omega)_{ij}$$

Рассмотрим следующее произведение

$$U_{\mathbf{w}}^T U_{\mathbf{w}} = U_{\mathbf{w}} U_{\mathbf{w}} = (I - 2\Omega)(I - 2\Omega) = I - 4\Omega + 4\Omega\Omega = I$$

Из ортогональности матрицы следует, что она сохраняет скалярное произведение. Зафиксируем собственный вектор $\mathbf{e}$:

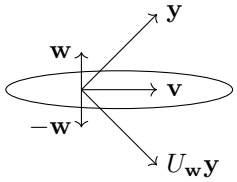
$$(U_{\mathbf{w}}\mathbf{e}, U_{\mathbf{w}}\mathbf{e})_2 = (\lambda\mathbf{e}, \lambda\mathbf{e})_2 = \lambda^2(\mathbf{e}, \mathbf{e}) \Rightarrow 1 = \lambda^2$$

2. $(\Omega\mathbf{w})_i = \left(\sum_{k=1}^{n} w_i w_k w_k\right)_i = \left(w_i \sum_{k=1}^{n} w_k^2\right)_i = w_i \Rightarrow (1 - 2\Omega)\mathbf{w} = \mathbf{w} - 2\mathbf{w} = -\mathbf{w}$

3. $(\Omega\mathbf{v})_i = \left(\sum_{k=1}^{n} w_i w_k v_k\right)_i = (w_i(\mathbf{w}, \mathbf{v})_2)_i = 0 \Rightarrow (1 - 2\Omega)\mathbf{v} = \mathbf{v} - 0 = \mathbf{v}$

4. $\forall\ \mathbf{y} \in \mathbb{R}^n : \mathbf{y} = \alpha\mathbf{v} + \beta\mathbf{w},\ \mathbf{w} \perp \mathbf{v} \Rightarrow$

   $\Rightarrow U_{\mathbf{w}}\mathbf{y} = \alpha U_{\mathbf{w}}\mathbf{v} + \beta U_{\mathbf{w}}\mathbf{w} = \alpha\mathbf{v} - \beta\mathbf{w}$



□

Преобразование Хаусхолдера

Ставится следующая задача: заданы два вектора $\mathbf{y}$ и $\mathbf{e}$ единичной длины. Требуется найти единичный вектор $\mathbf{w}$ такой, что $U_{\mathbf{w}}\mathbf{y} = \mathbf{e}$.

Для решения требуется воспользоваться тем, что образ оператора $U_{\mathbf{w}}\mathbf{y}$, $\mathbf{y} \in R^n$, является зеркальным отображением относительно гиперплоскости, перпендикулярной $\mathbf{w}$. Тогда искомый вектор можно представить как $\mathbf{w} = \pm(\mathbf{y} - \mathbf{e})/\|\mathbf{y} - \mathbf{e}\|_2$. Действительно,

$$(U_{\mathbf{w}}\mathbf{y})_i = ((I - 2\Omega)\mathbf{y})_i = (\mathbf{y} - 2\mathbf{w}\mathbf{w}^T\mathbf{y})_i = y_i - 2\xi_i$$

$$\xi_i = \frac{\sum_{k=1}^{n}(\mathbf{y} - \mathbf{e})_i(\mathbf{y} - \mathbf{e})_k y_k}{(\mathbf{y} - \mathbf{e}, \mathbf{y} - \mathbf{e})} = \frac{(y_i - e_i)\sum_{k=1}^{n}(y_k - e_k)y_k}{(\mathbf{y}, \mathbf{y}) - 2(\mathbf{y}, \mathbf{e}) + (\mathbf{e}, \mathbf{e})} = \frac{(y_i - e_i)\left(\sum_{k=1}^{n} y_k^2 - \sum_{k=1}^{n} e_k y_k\right)}{2 - 2(\mathbf{y}, \mathbf{e})} =$$

$$= \frac{(y_i - e_i)\,(1 - (\mathbf{y}, \mathbf{e}))}{2(1 - (\mathbf{y}, \mathbf{e}))} = \frac{y_i - e_i}{2} \Rightarrow (U_{\mathbf{w}}\mathbf{y})_i = y_i - 2\xi_i = y_i - 2\frac{y_i - e_i}{2} = y_i - y_i + e_i = e_i$$

Отметим, что преобразование $U_{\mathbf{w}}$ не меняет длины вектора (т.к. матрица ортогональна), следовательно, для неединичного вектора $\mathbf{y}$ имеем:

$$U_{\mathbf{w}}\mathbf{y} = \alpha\mathbf{e},\ \|\mathbf{y}\|_2 = \alpha \Rightarrow \mathbf{w} = \pm\frac{(\alpha^{-1}\mathbf{y} - \mathbf{e})}{\|\alpha^{-1}\mathbf{y} - \mathbf{e}\|_2} = \pm\frac{(\mathbf{y} - \alpha\mathbf{e})}{\|\mathbf{y} - \alpha\mathbf{e}\|_2}$$

Метод отражений

Произвольная квадратная матрица A может быть приведена к верхнетреугольному виду в результате последовательного умножения слева на ортогональные матрицы отражений. Действительно, по векторам $\mathbf{y}_1 = (a_{11}, \dots, a_{n1})^T$ и $\mathbf{e}_1 = (1, 0 \dots, 0)^T$ можно построить вектор $\mathbf{w}_1$ и матрицу $U_{\mathbf{w}_1}$ так, чтобы первый столбец матрицы $A^{(1)} = U_{\mathbf{w}_1} A$ был пропорционален вектору $\mathbf{e}_1 \in \mathbb{R}^n$, то есть $U_{\mathbf{w}_1}\mathbf{y}_1 = \pm\alpha_1\mathbf{e}_1$.

$$\mathbf{w}_1 = \left(\frac{a_{11}}{\alpha_1} + \operatorname{sgn}(a_{11}), \frac{a_{21}}{\alpha_1}, \dots, \frac{a_{n1}}{\alpha_1}\right)^T,\ \alpha_1 = \sqrt{\sum_{i=1}^{n} a_{i1}^2};\ \mathbf{w}_1 := \frac{\mathbf{w}_1}{\|\mathbf{w_1}\|_2}$$

Такой выбор знака и предварительная нормировка на α_1 гарантируют малость вычислительной погрешности и устойчивость алгоритма.

Далее в пространстве $\mathbb{R}^{n-1}$ по вектору $\mathbf{y}_2 = (a_{22}^{(1)}, \dots, a_{n2}^{(1)})^T$ строится матрица $U_{\mathbf{w}_2}$, отображающая его в вектор, коллинеарный $\mathbf{e} = (1, 0, \dots, 0) \in \mathbb{R}^{n-1}$. Затем определяется $U_{\mathbf{w}_2} = \begin{pmatrix} 1 & 0 \\ 0 & U'_{\mathbf{w}_2} \end{pmatrix}$ и рассматривается матрица $A^{(2)} = U_{\mathbf{w}_2} U_{\mathbf{w}_1} A$, и так далее. На k-м шаге имеем $U_{\mathbf{w}_k} = \begin{pmatrix} I_{k-1} & 0 \\ 0 & U'_{\mathbf{w}_k} \end{pmatrix}$. Таким образом, матрица отражений $U_{\mathbf{w}_k}$ строится по вектору $\mathbf{w}_k \in \mathbb{R}^n$, заданному следующим образом

$$\mathbf{w}_k = \left(0, \dots, 0, \frac{a_{kk}^{(k-1)}}{\alpha_k} + \operatorname{sgn}(a_{kk}^{(k-1)}), \frac{a_{k+1,k}^{(k-1)}}{\alpha_k}, \dots, \frac{a_{nk}^{(k-1)}}{\alpha_k}\right)^T,\ \alpha_k = \sqrt{\sum_{i=k}^{n} (a_{ik}^{(k-1)})^2};\ \mathbf{w}_k = \frac{\mathbf{w}_k}{\|\mathbf{w_k}\|_2}$$

В результате преобразований получится верхняя треугольная матрица $R = UA$. При практической реализации явное вычисление $U_{\mathbf{w}_k}$ не требуется, так как $U_{\mathbf{w}_k} A^{(k-1)} = A^{(k-1)} - 2\mathbf{w}_k(\mathbf{w}_k^T A^{(k-1)})$. При этом изменяются только элементы $a_{ij}^{(k-1)}, k \le i, j \le n$ матрицы $A^{(k-1)}$.

Из условия $UU = I$ имеем $A = UR$. Таким образом, произвольная квадратная матрица A может быть представлена в виде произведения симметричной ортогональной матрицы U и верхней треугольной матрицы R.

Рассмотренный алгоритм позволяет свести СЛАУ $A\mathbf{x} = \mathbf{b}$ к виду $R\mathbf{x} = U_{\mathbf{w}_{n-1}} \dots U_{\mathbf{w}_1}\mathbf{b}$, а затем найти ее решение обратным ходом метода Гаусса. Пусть решается задача с возмущенной правой частью $A\mathbf{x} = \mathbf{b} + \delta,\ \|\delta\|_2 \ll \|\mathbf{b}\|$. Так как ортогональные преобразования не меняют евклидову норму векторов, то для приведенной системы $R\mathbf{x} = U\mathbf{b} + U\delta$ имеем $\|\mathbf{U}\delta\|_2 = \|\delta\|_2 \ll \|\mathbf{b}\|_2 = \|\mathbf{U}\mathbf{b}\|_2$, и относительная погрешность правой части не увеличилась. В методе Гаусса матрица преобразования C не ортогональна, и в общем случае может оказаться, что $\|\mathbf{C}\delta\|_2$ станет сравнимым с $\|\mathbf{C}\mathbf{b}\|_2$. То есть малая начальная погрешность может существенно исказить ответ.

Пример 25.1.

$$A^{(0)} = \begin{pmatrix} 1.00 & 2.00 & 3.00 \\ 3.00 & 1.00 & 2.00 \\ 2.00 & 3.00 & 1.00 \end{pmatrix};\ \mathbf{b}_0 = \begin{pmatrix} 3.00 \\ 6.00 \\ 9.00 \end{pmatrix};\ \mathbf{w}_0 = \begin{pmatrix} 0.80 \\ 0.50 \\ 0.34 \end{pmatrix};\ U_{\mathbf{w}_0} = \begin{pmatrix} -0.27 & -0.80 & -0.53 \\ -0.80 & 0.49 & -0.34 \\ -0.53 & -0.34 & 0.77 \end{pmatrix}$$

$$A^{(1)} = \begin{pmatrix} -3.74 & -2.94 & -2.94 \\ 0.00 & -2.13 & -1.76 \\ 0.00 & 0.92 & -1.51 \end{pmatrix};\ \mathbf{b}_1 = \begin{pmatrix} -10.42 \\ -2.49 \\ 3.34 \end{pmatrix};\ \mathbf{w}_1 = \begin{pmatrix} 0.00 \\ -0.98 \\ 0.20 \end{pmatrix};\ U_{\mathbf{w}_1} = \begin{pmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & -0.92 & 0.40 \\ 0.00 & 0.40 & 0.92 \end{pmatrix}$$

$$A^{(2)} = \begin{pmatrix} -3.74 & -2.94 & -2.94 \\ 0.00 & 2.31 & 1.02 \\ 0.00 & 0.00 & -2.08 \end{pmatrix};\ \mathbf{b}_2 = \begin{pmatrix} -10.42 \\ 3.61 \\ 2.08 \end{pmatrix};\ \mathbf{w}_2 = \begin{pmatrix} 0.00 \\ 0.00 \\ -1.00 \end{pmatrix};\ U_{\mathbf{w}_2} = \begin{pmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & -1.00 \end{pmatrix}$$

После окончания метода отражений входные матрица и вектор имеют следующий вид:

$$A^{(3)} = \begin{pmatrix} -3.74 & -2.94 & -2.94 \\ 0.00 & 2.31 & 1.02 \\ 0.00 & 0.00 & 2.08 \end{pmatrix};\ \mathbf{b}_3 = \begin{pmatrix} -10.42 \\ 3.61 \\ -2.08 \end{pmatrix}$$

После применения обратного хода Гаусса получили итоговый ответ

$$A = \begin{pmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 1.00 \end{pmatrix};\ \mathbf{x} = \begin{pmatrix} 2.00 \\ 2.00 \\ -1.00 \end{pmatrix}$$

26 Невырожденная ЗНК: метод нормального уравнения, метод QR-разложения.

Пусть требуется решить СЛАУ $A\mathbf{x} = \mathbf{b}$ с матрицей A размерности $m \times n$, правой частью $\mathbf{b} \in \mathbb{R}^m$ и вектором неизвестных $x \in \mathbb{R}^n$. Рассмотрим три случая:

1. $m = n$, $\det(A) \neq 0$. Задача невырождена и вектор $\mathbf{x} = A^{-1}\mathbf{b}$ является точным решением. Для вектора невязки $\mathbf{r} = \mathbf{b} - A\mathbf{x}$ имеем $\|\mathbf{r}\| = 0$.

$$n\begin{pmatrix} & A & \\ & & \\ & n & \end{pmatrix}\begin{pmatrix} & \mathbf{x} & \\ & & \\ & n & \end{pmatrix} = \begin{pmatrix} & \mathbf{b} & \\ & & \\ & n & \end{pmatrix}$$

2. $m < n$, $\mathrm{rk}(A) = m$. Задача недоопределена. Исходная система имеет подпространство решений размерности $n - m$.

$$m\begin{pmatrix} & A & \\ & & \\ & n & \end{pmatrix}\begin{pmatrix} & \mathbf{x} & \\ & & \\ & n & \end{pmatrix} = \begin{pmatrix} & \mathbf{b} & \\ & & \\ & m & \end{pmatrix}$$

3. $m > n$, $\mathrm{rk}(A) = n$. Система переопределена и, если она несовместна, то точного решения не существует, т.е. для произвольного $\mathbf{x} \in \mathbb{R}^n$ имеем $\|\mathbf{b} - A\mathbf{x}\| = \|r\| > 0$.

$$m\begin{pmatrix} & A & \\ & & \\ & & \\ & n & \end{pmatrix}\begin{pmatrix} & \mathbf{x} & \\ & & \\ & n & \end{pmatrix} = \begin{pmatrix} & \mathbf{b} & \\ & & \\ & & \\ & m & \end{pmatrix}$$

Представляют интерес методы решения переопределенных задач, поэтому далее, если не оговаривается иное, считаем, что $m > n$, $\mathrm{rk}(A) = n$. Для задач такого рода Гаусс предложил считать решением вектор $\mathbf{x}$, минимизирующий евклидову норму вектора невязки $\inf \mathbf{y} \|\mathbf{b} - A\mathbf{y}\|_2 = \|\mathbf{b} - A\mathbf{x}\|_2$.

Рассмотрим некоторые методы решения данной минимизационной задачи, называемой задачей наименьших квадратов (ЗНК).

Метод нормального уравнения

Рассмотрим следующую, называемую нормальной, систему уравнений $A^TA\mathbf{x} = A^T\mathbf{b}$ с квадратной матрицей $A^TA \in \mathbb{R}^{n\times n}$. Отсюда найдем вектор $\mathbf{x}$.

Теорема 26.1 (Гаусс К.Ф.)**.** *Пусть $m \geq n$, $rk(A) = n$. Тогда нормальное уравнение имеет единственное решение.*

Доказательство. Рассмотрим $B = A^TA$, тогда $B = B^T$, так как $B^T = (A^TA)^T = A^TA = B$ и

$$(B\mathbf{x}, \mathbf{x}) = (A^TA\mathbf{x}, \mathbf{x}) = (A\mathbf{x}, A\mathbf{x}) > 0,\ \forall \mathbf{x} \neq 0$$

Значит, B является положительно определенной, значит невырожденной, значит $\exists!$ решение задачи. □

Покажем, что решение нормального уравнения является решением исходной задачи

Теорема 26.2. *Пусть $m \geq n$, $rk(A) = n$. Вектор $\mathbf{x}$ является решением ЗНК $\inf_{\mathbf{y}} \|\mathbf{b} - A\mathbf{y}\|_2$ тогда и только тогда, когда $\mathbf{x}$ является решением нормального уравнения $A^TA\mathbf{x} = A^T\mathbf{b}$.*

Доказательство. Из предыдущей теоремы известно, что $\exists!$ $\mathbf{x}$ - решение нормального уравнения, тогда

$$A^TA\mathbf{x} - A^T\mathbf{b} = 0 \Leftrightarrow A^T(A\mathbf{x} - \mathbf{b}) = 0$$

Возьмем $\mathbf{y} = \mathbf{x} + \boldsymbol{\Delta}$, тогда

$$\|A\mathbf{y} - \mathbf{b}\|_2^2 = (A\mathbf{x} - \mathbf{b} + A\boldsymbol{\Delta}, A\mathbf{x} - \mathbf{b} + A\boldsymbol{\Delta})_2 = \|A\mathbf{x} - \mathbf{b}\|_2^2 + 2(\Delta, \underbrace{A^T(A\mathbf{x} - \mathbf{b})}_{=0})_2 + \|\boldsymbol{\Delta}\|_2^2$$

Следовательно, минимум достигается при $\boldsymbol{\Delta} = 0$, то есть на векторе $\mathbf{y} = \mathbf{x}$. □

Метод нормального уравнения прост в реализации, однако в приближенной арифметике для почти вырожденных задач большой размерности может давать плохой результат. Например, в случае квадратной матрицы $A^T = A$ имеем $\mathrm{cond}_2(A^TA) = \mathrm{cond}_2(A)^2$. Таким образом, обусловленность исходной задачи возводится в квадрат, поэтому полученное численно решение может сильно отличаться от точного, если $\mathrm{cond}(A) \gg 1$. Рассмотрим более устойчивый метод.

QR-разложение

Теорема 26.3. *Пусть A - $m \times n$ матрица, $m \geq n$, $rk(A) = n$. Тогда $\exists!$ матрица $Q \in \mathbb{R}^{m\times n}$, $Q^TQ = I_n$, и $R \in \mathbb{R}^{n\times n}$, $R_{ii} > 0$ - верхнетреугольная такие, что $A = QR$.*

Доказательство. С помощью алгоритма ортогонализации Грама-Шмидта по набору $< \mathbf{a}_1, \ldots, \mathbf{a}_n >$ из столбцов матрицы A построим ортонормальный базис $< \mathbf{q}_1, \ldots, \mathbf{q}_n >$. По найденным векторам построим матрицу $Q = (\mathbf{q}_1, \ldots, \mathbf{q}_n)$ и рассмотрим $R = Q^TA$.

$$Q^TA = \begin{pmatrix} \mathbf{q}_1^T \\ \vdots \\ \mathbf{q}_n^T \end{pmatrix} (\mathbf{a}_1, \ldots, \mathbf{a}_n) = \begin{pmatrix} (\mathbf{q}_1, \mathbf{a}_1) & (\mathbf{q}_1, \mathbf{a}_2) & & (\mathbf{q}_1, \mathbf{q}_n) \\ 0 & (\mathbf{q}_2, \mathbf{a}_2) & & (\mathbf{q}_2, \mathbf{q}_n) \\ \ldots & \ldots & \ddots & \ldots \\ 0 & 0 & & (\mathbf{q}_n, \mathbf{q}_n) \end{pmatrix} = R$$

Обратим внимание, что $(\mathbf{q}_i, \mathbf{a}_j) = 0$ при $i > j$ по построению базиса через алгоритм Грама-Шмидта. Так же построенная матрица Q является ортогональной, так как $(\mathbf{q}_i, \mathbf{q}_j) = \delta_{ij}$. Отсюда и из $Q^TA = R$ имеем $A = QR$.

Единственность в данном случае следует из факта, что построить соответствующее линейное подпространство натянутое на $\{\mathbf{q}_i\}_i$ можно единственным образом, если договориться, что элементы на диагонали матрицы R положительны. Тогда каждый следующий вектор из базиса будет иметь направление совпадающее с направлением, полученным по правилу правой руки. □

Замечание 26.1. Если модифицированный алгоритм Грама–Шмидта необходимо применить для решения задачи $A\mathbf{x} = \mathbf{b}$, то задачу сводят к системе $R\mathbf{x} = \mathbf{d}$. При этом матрица R находится указанным способом, а компоненты вектора $\mathbf{d}$ для сохранения вычислительной устойчивости определяются по следующему алгоритму: $\mathbf{r} = \mathbf{b}$, $d_i = (\mathbf{r}_{i-1}, \mathbf{q}_i)$, $\mathbf{r}_i = \mathbf{r}_{i-1} - d_i\mathbf{q}_i$, $i = 1, \ldots, n$. i i

Замечание 26.2. Отметим, что если приближенное решение $\mathbf{x}$ системы $A\mathbf{x} = \mathbf{b}$ получено каким-либо методом, основанном на QR-разложении, то можно выполнить следующий процесс уточнения. Найдем вектор невязки $\mathbf{r} = \mathbf{b} - A\tilde{\mathbf{x}}$ с удвоенным количеством значащих цифр и решим систему $A\mathbf{z} = \mathbf{r}/\|\mathbf{r}\|$. Положим $\tilde{\mathbf{x}} = \tilde{\mathbf{x}} + \|\mathbf{r}\|\,\mathbf{z}$. Процесс уточнения значительно экономичнее, чем решение исходного уравнения, так как разложение матрицы A уже имеется. Уточнение можно повторять до тех пор, пока убывает норма вектора невязки.

Подобное разложение $A = QR$ можно построить и с квадратной матрицей $Q \in \mathbb{R}^{m\times m}$.

Теорема 26.4. *Пусть $m > n$, $rk(A) = n$. Тогда матрицу $A \in \mathbb{R}^{m\times n}$ можно привести к виду*

$$A = QR = \left(\begin{array}{c|c} Q_1 & Q_2 \\ m \times n & m \times (m-n) \end{array} \right) \left(\begin{array}{c} R_1 \\ \hline 0 \\ (m-n) \times n \end{array} \right) = Q_1R_1$$

$$Q \in \mathbb{R}^{m\times m},\ Q^TQ = I,\ R \in \mathbb{R}^{m\times n},\ \det R_1 \neq 0$$

$R_1 \in \mathbb{R}^{n\times n}$ - верхнетреугольная. В этом случае решение $\mathbf{x}$ задачи $R_1\mathbf{x} = Q_1^T\mathbf{b}$ является решением ЗНК.

Доказательство. Матрицу $Q_1 \in \mathbb{R}^{m\times n}$ мы строим аналогично предыдущей теореме. Можем для этого использовать или алгоритм Грама-Шмидта, или метод отражений, или метод вращений. Для выполнения требования $Q^TQ = I$ в Q_2 мы поставим дополнительный базис к Q_1 из пространства $\mathbb{R}^m$, то есть $(m-n)$ векторов, ортонормальные векторам из Q_1.

Покажем, что решение $R_1\mathbf{x} = Q_1^T\mathbf{b}$ является решением ЗНК. Действительно, так как решение нормального уравнения является решением ЗНК, то

$$A^TA\mathbf{x} = A^T\mathbf{b} \Leftrightarrow (QR)^T(QR)\mathbf{x} = (QR)^T\mathbf{b} \Leftrightarrow R^TR\mathbf{x} = R^TQ^T\mathbf{b} \Leftrightarrow R\mathbf{x} = Q^T\mathbf{b} \Leftrightarrow R_1\mathbf{x} = Q_1^T\mathbf{b}$$

□

27 Задача наименьших квадратов неполного ранга: методы QR-разложения и QR-разложения с выбором главного столбца

Методы QR-разложения

ЗНК называется вырожденной, если $\mathrm{rk}(A) < n$, $\det(A^TA) = 0$. При численном решении вырожденных и почти вырожденных систем требуется изменить постановку задачи и соответственно применять иные методы.

Замечание 27.1. Рассмотрим следующий пример

$$\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

В данном уравнении $m = n = 2$ и задача вырождена, можно выбрать семейство решений $\mathbf{x} = (1 - x_2, x_2)^T$, $\mathbf{r} = \mathbf{b} - A\mathbf{x} = (0, 1)^T$. Можно выбрать решения как с нормами порядка единицы, так и со сколь угодно большими. Однако, для возмущенной задачи

$$\begin{pmatrix} 1 & 1 \\ 0 & \varepsilon \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

формально близкой к исходной при $\varepsilon \approx 0$, имеется единственное решение $(1 - \varepsilon^{-1}, \varepsilon^{-1})^T$ с большой нормой порядка ε^{-1} и нулевой невязкой. Это означает, что сколь угодно малое возмущение элементов матрицы может существенно изменить структуру и норму решения.

Теорема 27.1. *Пусть дана $A \in \mathbb{R}^{m\times n}$, $rk(A) = k < n \le m$. Тогда множество векторов - решений ЗНК - образует линейное подпространство размерности $n - k$.*

Доказательство. Так как $\mathrm{rk}(A) = k < n$, то $\dim(\ker(A)) = n - k$. Рассмотрим $\mathbf{y} \in \ker(A)$. Утверждается, если $\mathbf{x}$ - решение ЗНК, то есть $\inf_{\mathbf{z}} \|\mathbf{b} - A\mathbf{z}\|_2 = \|\mathbf{b} - A\mathbf{x}\|$, то $\mathbf{x}+\mathbf{y}$ является решением ЗНК. Действительно

$$\|\mathbf{b} - A(\mathbf{x} + \mathbf{y})\|_2 = \|\mathbf{b} - A\mathbf{x} - A\mathbf{y}\|_2 = \|\mathbf{b} - A\mathbf{x}\|_2$$

□

Теорема 27.2. *Пусть дана $A \in \mathbb{R}^{m\times n}$, $rk(A) = k < n \le m$. Тогда матрицу можно привести к следующему виду*

$$A = QR = \left(\begin{array}{c|c} Q_1 & Q_2 \\ m \times k & m \times (m-k) \end{array} \right) \left(\begin{array}{c|c} R_1 & R_2 \\ k \times k & k \times (n-k) \\ \hline 0 & 0 \\ (m-k) \times k & (m-k) \times (n-k) \end{array} \right)$$

$Q \in \mathbb{R}^{m\times m}$, $Q^TQ = I$, R_1 - верхнетреугольная, $\det(R_1) \neq 0$. Для ЗНК с матрицей A имеется семейство решений

$$\mathbf{x} = (R_{11}^{-1}(Q_1\mathbf{b} - R_{12}\mathbf{x_2}), \mathbf{x}_2)^T,\ \mathbf{x}_1 \in \mathbb{R}^k,\ \mathbf{x}_2 \in \mathbb{R}^{n-k}$$

Доказательство. Для того, чтобы понять почему матрицы имеют такой вид, построим матрицу Q размера $m \times m$ как ортонормальные вектора из $\mathbb{R}^m$ с помощью алгоритма Грама-Шмидта. Посмотрим на произведение следующих матриц

$$Q^TA = \begin{pmatrix} \mathbf{q}_1 \\ \vdots \\ \mathbf{q}_k \\ \mathbf{q}_{k+1} \\ \vdots \\ \mathbf{q}_m \end{pmatrix} \Big(\mathbf{a}_1, \ldots, \mathbf{a}_k, \mathbf{a}_{k+1}, \ldots, \mathbf{a}_n\Big) = \left(\begin{array}{c|c} R_{11} & R_{12} \\ k \times k & k \times (n-k) \\ \hline R_{21} & R_{22} \\ (m-k) \times k & (m-k) \times (n-k) \end{array} \right)$$

1. $R_{11} = \begin{pmatrix} (\mathbf{q}_1, \mathbf{a}_1) & & (\mathbf{q}_1, \mathbf{a}_k) \\ & \ddots & \\ (\mathbf{q}_k, \mathbf{a}_1) & & (\mathbf{q}_k, \mathbf{a}_k) \end{pmatrix} = \begin{pmatrix} (\mathbf{q}_1, \mathbf{a}_1) & & (\mathbf{q}_1, \mathbf{a}_k) \\ & \ddots & \\ 0 & & (\mathbf{q}_k, \mathbf{a}_k) \end{pmatrix}$ и $R_{21} = 0$ так как $(\mathbf{q}_i, \mathbf{a}_j) = 0,\ i > j$

2. Рассмотрим произвольный элемент матрицы R_{12}: $(\mathbf{q}_i, \mathbf{a}_j)$, $1 \le i \le k$, $k < j \le n$. Так как $\mathrm{rk}(A) = k$, то $\mathbf{a}_j$ можно выразить через $\mathbf{q}_t$, $1 \le t \le k$: $\left(\mathbf{q}_i, \sum_{t=1}^{k} c_t \mathbf{q}_t\right) = c_i$. То есть матрица R_{12} состоит из компонент векторов разложенных по $\mathbf{q}_i$, $1 \le i \le k$. И так как в разложении этих векторов не участвуют $\mathbf{q}_j$, $k < j \le n$, то соответствующие коэффициенты в разложении равны 0, а значит $R_{22} = 0$.

Решим задачу наименьших квадратов. Исходную задачу домножим на Q^T. Итоговая норма не изменится, так как Q^T является ортогональной матрицей, то есть сохраняет длину.

$$\inf_{\tilde{\mathbf{x}}} \|\mathbf{b} - A\tilde{\mathbf{x}}\|_2^2 = \inf_{\tilde{\mathbf{x}}} \left\|Q^T\mathbf{b} - R\tilde{\mathbf{x}}\right\|_2^2 = \inf_{\tilde{\mathbf{x}}} \left\| \begin{pmatrix} Q_1^T \\ Q_2^T \end{pmatrix} \mathbf{b} - \begin{pmatrix} R_{11} & R_{12} \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \tilde{\mathbf{x}}_1 \\ \tilde{\mathbf{x}}_2 \end{pmatrix} \right\|_2^2$$

$$= \inf_{\tilde{\mathbf{x}}} \left\| \begin{matrix} Q_1^T\mathbf{b} - R_{11}\tilde{\mathbf{x}}_1 - R_{12}\tilde{\mathbf{x}}_2 \\ Q_2^T\mathbf{b} \end{matrix} \right\|_2^2 = \inf_{\tilde{\mathbf{x}}} \left(\left\|Q_1^T\mathbf{b} - R_{11}\tilde{\mathbf{x}}_1 - R_{12}\tilde{\mathbf{x}}_2\right\|_2^2 + \left\|Q_2^T\mathbf{b}\right\|_2^2 \right)$$

Минимум будет достигаться только если нам удастся занулить первое слагаемое. Тогда решение можно молучить из равенства

$$\left\|Q_1^T\mathbf{b} - R_{11}\mathbf{x}_1 - R_{12}\mathbf{x}_2\right\|_2^2 = 0 \Leftrightarrow \mathbf{x}_1 = R_{11}^{-1}(Q_1^T\mathbf{b} + R_{12}\mathbf{x}_2),\ \forall\ \mathbf{x}_2 \in \mathbb{R}^{n-k}$$

Таким образом, пространство решений имеет размер $n - k$. □

Замечание 27.2. Научились решать задачу $\inf_{\tilde{\mathbf{x}}} \|\mathbf{b} - A\tilde{\mathbf{x}}\|_2$, но что если от нас требуется найти $\tilde{\mathbf{x}}_2$: $\|\mathbf{x}\| \to \inf$? Обычно выбор $\tilde{\mathbf{x}_2} \equiv 0$ может давать неплохое приближение, но в общем случае бывают задачи, где такой ответ не подходит.

QR-разложение с выбором главного столбца

Преобразуем исходную задачу так, чтобы первые k столбцов полученной матрицы A были линейно независимы. Перестановки столбцов в матрице A удобно проводить в процессе вычислений. Цель соответствующих перестановок - получить в матрице $R = \begin{pmatrix} R_{11} & R_{12} \\ 0 & R_{22} \\ 0 & 0 \end{pmatrix}$ как можно лучше обусловленный блок R_{11} и как можно меньшие элементы в R_{22}. В машинной арифметике принято считать, что блок R_{22} обнулился, если все вектора в нем размера меньше фиксированного машинного ε.

Соответствующие вычисления проводятся на основе стандартного QR- разложения, и так как удобнее проводить алгоритм итеративно, то чаще используется метод отражений. На k-м шаге, $1 \le k \le n-1$, в матрице A выбирают столбец с номером j_k, $k \le j_k \le n$, с наибольшей величиной $\max_{k \le j \le n} \left\|\mathbf{a}_j^{(k)}\right\|_2$: Если таких столбцов несколько, то берут произвольный из них. В матрице A найденный столбец j переставляют с k-м столбцом. Далее реализуют очередной шаг QR-разложения.

В конце алгоритма переставляют компоненты решения x_i в соотвествии с соответствии с перестановками столбцов.

Замечание 27.3. Для невырожденной задачи $A\mathbf{x} = \mathbf{b}$ с матрицей $A \in \mathbb{R}^{n \times n}$ вида

$$A = \begin{pmatrix} 1 & -1 & \dots & & -1 \\ 0 & 1 & & & \\ & & \ddots & & \vdots \\ & & & 1 & -1 \\ 0 & & & & 1 \end{pmatrix}$$

QR-алгоритм с выбором максимального элемента даст результат намного хуже, чем стандартный. При стандартном QR ответ записывается сразу: $Q \equiv I$, $R \equiv A$. Тогда как при выборе максимального последний элемент R_{nn} будет иметь порядок $\underline{\underline{\mathcal{O}}}(2^{-n})$. То есть, если мы возьмем $n = 50$, то последний элемент будет сравним с машинным нулем, и алгоритм вполне может посчитать получившуюся матрицу вырожденной, хотя задача очевидно таковой не является.

28 Сингулярное разложение.

Утверждение. *Пусть* $A \in \mathbb{R}^{m\times n}$, $m \geq n$. *Тогда справедливо сингулярное разложение*

$$A = U\Sigma V^T = (U_1 U_2)\begin{pmatrix}\Sigma_n \\ 0\end{pmatrix} V^T$$

- $U \in \mathbb{R}^{m\times m} = (\mathbf{u}_1, \ldots, \mathbf{u}_m)$ - *ортогональная матрица левых сингулярных векторов,* $U_1 \in \mathbb{R}^{m\times n}$, $U_2 \in \mathbb{R}^{m\times(m-n)}$
- $V \in \mathbb{R}^{n\times n} = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$ - *ортогональная матрица правых сингулярных векторов.*
- $\Sigma_n \in \mathbb{R}^{n\times n}$ - *диагональная матрица, на диагонали которой расположены упорядоченные сингулярные числа* $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_n \geq 0$.

Если $m = n$, то $\Sigma = \Sigma_n$.

Построив SVD-разложение можно установить, является ли задача вырожденной ($\sigma_n = 0$), невырожденной ($\sigma_n \neq 0$), 'хорошей' (σ_1/σ_n не слишком велико).

Если $m < n$; то сингулярное разложение строят для матрицы A^T. Если $A = A^T$, то сингулярные числа $\sigma_i = |\lambda_i(A)|$, т.е. с точностью до знака совпадают с собственными числами, сингулярные векторы $\mathbf{v}_i$ являются соответствующими собственными векторами. В случае, если $\lambda_i(A) < 0$, то минус утаскивают в соответствующий вектор в U, и поэтому разложение возможно.

Как найти U и V? Рассмотрим следующие выражения

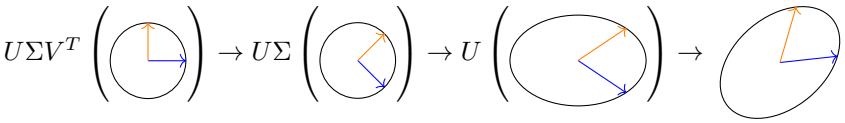
$$AA^T = U\Sigma V^T (U\Sigma V^T)^T = U\Sigma V^T V^T \Sigma^T U^T = U\Sigma_n^2 U^T$$

$$A^T A = (U\Sigma V^T)^T U\Sigma V^T = V\Sigma U^T U\Sigma^T V^T = V\Sigma_n^2 V^T$$

Таким образом, U и V состоят из собственных векторов для матрицы $AA^T \in \mathbb{R}^{m\times m}$ и $A^T A \in \mathbb{R}^{n\times n}$ соответственно, которые для данных симметричных матриц являются ортогональными. Такие вектора можно найти с помощью разложения матрицы в Жорданову форму.

Геометрическая интерпретация SVD-разложения.

Рассмотрим оператор A, переводящий элемент $\mathbf{x} \in \mathbb{R}^n$ в элемент $\mathbf{y} \in \mathbb{R}^m$. Единичная сфера из $\mathbb{R}^n$ под действием A переходит в эллипсоид в подпространстве span $< \mathbf{u}_1, \ldots, \mathbf{u}_n > \subset \mathbb{R}^m$. Вектора $\mathbf{u}_i$ задают полуоси эллипсоида, $\mathbf{v}_i$ их прообразы, σ_i коэффициенты удлинения векторов $\mathbf{v}_i$.



Алгебраическая интерпретация SVD-разложения. Рассмотрим оператор A, переводящий элемент $\mathbf{x} \in \mathbb{R}^n$ в элемент $\mathbf{y} \in \mathbb{R}^m$. В пространстве $\mathbb{R}^n$ $\exists$ ортонормальный базис $\mathbf{v}_i$, а в пространстве $\mathbb{R}^m$ - ортонормальный базис $\mathbf{v}_j$. Тогда для произвольного $\mathbf{x} \in \mathbb{R}^n$:

$$A\mathbf{x} = U\Sigma V^T \mathbf{x} = U\Sigma V^T \sum_{i=1}^{n} \alpha_i \mathbf{v}_i = U\Sigma \sum_{i=1}^{n} \alpha_i \mathbf{e}_i = U \sum_{i=1}^{n} \sigma_i \alpha_i \mathbf{e}_i = \sum_{i=1}^{n} \sigma_i \alpha_i \mathbf{u}_i$$

А вектора $\mathbf{u}_{n+1}, \ldots, \mathbf{u}_m$ добавляются как ортонормальное добавление к векторам $\mathbf{u}_1, \ldots, \mathbf{u}_n$ для соблюдения $U^T U = I$.

29 Решение задачи наименьших квадратов полного и неполного рангов методом сингулярного разложения.

Опр. 29.1. Пусть $A \in \mathbb{R}^{m\times n}$, $m \geq n$. Тогда справедливо *сингулярное* разложение

$$A = U\Sigma V^T = (U_1U_2)\begin{pmatrix} \Sigma_n \\ 0 \end{pmatrix} V^T$$

- $U \in \mathbb{R}^{m\times m} = (\mathbf{u}_1, \ldots, \mathbf{u}_m)$ - ортогональная матрица *левых сингулярных векторов*, $U_1 \in \mathbb{R}^{m\times n}$, $U_2 \in \mathbb{R}^{m\times(m-n)}$
- $V \in \mathbb{R}^{n\times n} = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$ - ортогональная матрица *правых сингулярных векторов.*
- $\Sigma_n \in \mathbb{R}^{n\times n}$ - диагональная матрица, на диагонали которой расположены упорядоченные *сингулярные числа* $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_n \geq 0$.

Теорема 29.1. *Пусть $A = U\Sigma V^T$, $rk(A) = n$. Тогда решением ЗНК является вектор*

$$\mathbf{x} = V\Sigma_n^{-1}U_1^T\mathbf{b}$$

Доказательство. Так как искомый вектор является решением нормального уравнения, то

$$A^TA\mathbf{x} = A^T\mathbf{b} \Leftrightarrow \mathbf{x} = (A^TA)^{-1}A^T\mathbf{b} \Leftrightarrow \mathbf{x} = (V\Sigma_n^2V^T)^{-1}V^T\Sigma^TU\mathbf{b} = V\Sigma_n^{-2}\Sigma_n^TU_1\mathbf{b} = V\Sigma_n^{-1}U_1\mathbf{b}$$

□

Теорема 29.2. *Пусть*

$$A = (U_1U_2)\begin{pmatrix} \Sigma_k & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix},\ rk(A) = k < n < m$$

Здесь $U_1 \in \mathbb{R}^{m\times k}$, $\Sigma_k \in \mathbb{R}^{k\times k}$, $V_1 \in \mathbb{R}^{n\times k}$. Тогда решением ЗНК является пространство решений размерности $n-k$ вида

$$\mathbf{x} = V_1\Sigma_k^{-1}U_1^T\mathbf{b} + V_2\mathbf{z},\ \mathbf{z} \in \mathbb{R}^{n-k}$$

Доказательство. Исходную задачу домножим на U^T. Итоговая норма не изменится, так как U^T является ортогональной матрицей, то есть сохраняет длину.

$$\inf_{\tilde{\mathbf{x}}} \|\mathbf{b} - A\tilde{\mathbf{x}}\|_2^2 = \inf_{\tilde{\mathbf{x}}} \left\|U^T\mathbf{b} - \Sigma V^T\tilde{\mathbf{x}}\right\|_2^2 = \inf_{\tilde{\mathbf{x}}} \left\|\begin{pmatrix} U_1^T \\ U_2^T \end{pmatrix}\mathbf{b} - \begin{pmatrix} \Sigma_k & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix}\begin{pmatrix} \tilde{\mathbf{x}}_1 \\ \tilde{\mathbf{x}}_2 \end{pmatrix}\right\|_2^2$$

$$= \inf_{\tilde{\mathbf{x}}} \left\| \begin{matrix} U_1^T\mathbf{b} - \Sigma_kV_1^T\tilde{\mathbf{x}}_1 \\ U_2^T\mathbf{b} \end{matrix} \right\|_2^2 = \inf_{\tilde{\mathbf{x}}} \left(\left\|U_1^T\mathbf{b} - \Sigma_kV_1^T\tilde{\mathbf{x}}_1\right\|_2^2 + \left\|U_2^T\mathbf{b}\right\|_2^2\right)$$

Таким образом, минимум достигается только при

$$U_1^T\mathbf{b} - \Sigma_kV_1^T\mathbf{x}_1 = 0 \Leftrightarrow \mathbf{x}_1 = V_1\Sigma_k^{-1}U_1^T\mathbf{b}$$

Найдем общий вид решения. Так как $\dim\ker(A) = n-k$ и $V_1^TV_2 = 0$ вследствие ортогональности $\mathbf{v}_i$, то $U_1^T\mathbf{b} - \Sigma_kV_1^T(\mathbf{x}_1 + V_2\mathbf{z}) = 0$. Значит общее решение имеет вид

$$\mathbf{x} = V_1\Sigma_k^{-1}U_1^T\mathbf{b} + V_2\mathbf{z},\ \mathbf{z} \in \mathbb{R}^{n-k}$$

□

Замечание 29.1. SVD-разложение в отличие от QR-разложения отвечает на вопрос какой вектор в пространстве решений $\mathbb{R}^{n-k}$ имеет минимальную норму. Действительно,

$$(\mathbf{x},\mathbf{x}) = (\mathbf{x}_1,\mathbf{x}_1) + 2(\mathbf{x}_1, V_2\mathbf{z}) + (V_2\mathbf{z}, V_2\mathbf{z}) = (\mathbf{x}_1,\mathbf{x}_1) + (\mathbf{z},\mathbf{z})$$

при $\mathbf{z} = 0$ достигается минимум.

30 ЗНК с линейными ограничениями–равенствами: методы исключения, обобщенного SVD, взвешиванием

Ставится задача

$$\begin{cases} A\mathbf{x} = \mathbf{b},\ A \in \mathbb{R}^{m\times n} & \text{в смысле ЗНК} \\ B\mathbf{x} = \mathbf{d},\ B \in \mathbb{R}^{p\times n} & \text{как СЛАУ} \end{cases}$$

То есть среди всех решений, удовлетворяющих $B\mathbf{x} = \mathbf{d}$, ищем такие $\mathbf{x}$, что $\inf_{\tilde{\mathbf{x}}} \|\mathbf{b} - A\tilde{\mathbf{x}}\|_2 = \|\mathbf{b} - A\mathbf{x}\|_2$. Другая запись имеет вид

$$\inf_{\tilde{\mathbf{x}}: B\mathbf{x}=\mathbf{d}} \|\mathbf{b} - A\tilde{\mathbf{x}}\|_2 \Leftrightarrow \inf_{\tilde{\mathbf{x}}: \|\mathbf{d}-B\mathbf{x}\|=0} \|\mathbf{b} - A\tilde{\mathbf{x}}\|_2$$

Если $\mathrm{rk}(B) = n = p$, то решение $\mathbf{x} = B^{-1}\mathbf{d}$ является решением исходной задачи.

Методы исключения

Пусть $\mathrm{rk}(B) = p < n$, то есть матрица $B \in \mathbb{R}^{p\times n}$ имеет полный строчный ранг. Мы умеем работать с полным столбцовым рангом, поэтому транспонируем эту матрицу. Для матрицы B^T, $\mathrm{rk}(B^T) = p$ построим QR-разложение.

$$B^T = QR = Q\begin{pmatrix} R_1 \\ 0 \end{pmatrix},\ R_1 \in \mathbb{R}^{p\times p},\ Q \in \mathbb{R}^{n\times n}$$

Тогда исходную задачу можно преобразовать с помощью замены переменных $Q^T\mathbf{x} = \mathbf{y}$, обоснованную тем, что исходя из знания Q мы всегда найдем $\mathbf{x}$.

$$\left\|\mathbf{d} - R^TQ^T\mathbf{x}\right\|_2 = \left\|\mathbf{d} - R^T\mathbf{y}\right\|_2 = \left\|\mathbf{d} - \begin{pmatrix} R_1^T & 0 \end{pmatrix}\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix}\right\|_2$$

Но R_1 мы знаем точно, поэтому $\mathbf{y}_1 = R_1^{-1}\mathbf{d}$. В итоге, с учетом $AQQ^T\mathbf{x} = (A_1A_2)\mathbf{y}$, $A_1 \in \mathbb{R}^{m\times p}$, $A_2 \in \mathbb{R}^{m\times(n-p)}$, исходная задача принимает вид

$$\inf_{\tilde{\mathbf{x}}: B\mathbf{x}=\mathbf{d}} \|\mathbf{b} - A\tilde{\mathbf{x}}\|_2 = \inf_{\tilde{\mathbf{x}}: B\mathbf{x}=\mathbf{d}} \left\|\mathbf{b} - AQQ^T\tilde{\mathbf{x}}\right\|_2 = \inf_{\substack{\tilde{\mathbf{y}}_1 = R_1^{-1}\mathbf{d} \\ \tilde{\mathbf{y}}_2}} \left\|\mathbf{b} - AQ\begin{pmatrix} \tilde{\mathbf{y}}_1 \\ \tilde{\mathbf{y}}_2 \end{pmatrix}\right\|_2 = \inf_{\tilde{\mathbf{y}}_2} \|(\mathbf{b} - A_1\mathbf{y}_1) - A_2\tilde{\mathbf{y}}_2\|_2$$

Таким образом, нам получилось свести исходную задачу к задаче наименьших квадратов, решив которую мы сможем восстановить искомый $\mathbf{x}$.

Замечание 30.1. $\mathrm{rk}(B) = r < p < n$, и первые r столбцов линейно независимы. Тогда

$$B^T = Q\begin{pmatrix} R_{11} & R_{12} \\ 0 & 0 \end{pmatrix};\ B\times\mathbf{x} = \begin{pmatrix} R_{11}^T & 0 \\ R_{12}^T & 0 \end{pmatrix}\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \end{pmatrix}$$

где $\mathbf{y}_1 \in \mathbb{R}^r$, $\mathbf{y}_2 \in \mathbb{R}^{n-r}$ переменные, полученные после замены. Если полученная система окажется несовместной, то решение исходной задачи не существует. Иначе исходная задача с ограничениями сводится к стандартной ЗНК относительно $\tilde{\mathbf{y}}_2$ при вычисленном из системы $R_{11}^T\mathbf{y}_1 = \mathbf{d}_1$ векторе $\mathbf{y}_1$

Метод обобщенного сингулярного разложения

Утверждение. *Пусть $A \in \mathbb{R}^{m\times n}$, $m \geq n$, $B \in \mathbb{R}^{p\times n}$, $p \leq n$. Тогда справедливо разложение*

$$A = UD_AX^{-1};\ B = VD_BX^{-1}$$

- $U \in \mathbb{R}^{m\times m}$ - *ортогональная матрица*
- $V \in \mathbb{R}^{p\times p}$ - *ортогональная матрица*
- $X \in \mathbb{R}^{n\times n}$ - *обратимая матрица*
-

$$D_A = \begin{pmatrix} \alpha_1 & & 0 \\ & \ddots & \\ & & \alpha_n \\ \cdots & & \\ 0 & \ldots & 0 \end{pmatrix} \in \mathbb{R}^{m\times n} \quad D_B = \begin{pmatrix} \beta_1 & & 0 & & 0 \\ & \ddots & & \ldots & \\ & & \beta_p & & 0 \end{pmatrix} \in \mathbb{R}^{p\times n}$$

$$\alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_n \geq 0 \qquad \beta_1 \geq \beta_2 \geq \ldots \geq \beta_p \geq 0$$

Пусть нам известно сингулярное разложение, то есть известны $\mathbf{u}_i$, $i = 1, \dots, m$, $\mathbf{v}_j$, $j = 1, \dots, p$ и $\mathbf{x}_k$, $k = 1, \dots, n$. Положим $\mathbf{y} = X^{-1}\mathbf{x}$. Тогда исходная задача принимает вид

$$\begin{cases} D_B\mathbf{y} = V^T\mathbf{d} \\ \arg\inf_{\tilde{\mathbf{y}}} \left\|D_A\tilde{\mathbf{y}} - U^T\mathbf{b}\right\|_2^2 \end{cases}$$

Из первого уравнения находим первые p координат $\mathbf{y}$, затем подставим точные значения во второе уравнение и получим точные значения, на которых достигается inf:

$$y_j = \frac{(\mathbf{v}_i, \mathbf{d})}{\beta_j},\ j = 1, \dots, p; \quad y_i = \frac{(\mathbf{u}_i, \mathbf{b})}{\alpha_i},\ i = p+1 \dots, n$$

Итоговый ответ будет иметь вид

$$\mathbf{x} = X\mathbf{y} = \sum_{j=1}^{p} \frac{(\mathbf{v}_j, \mathbf{d})}{\beta_j}\mathbf{x}_j + \sum_{i=p+1}^{n} \frac{(\mathbf{u}_i, \mathbf{b})}{\alpha_i}\mathbf{x}_i$$

Замечание 30.2. При $\mathrm{rk}(B) = r < p$ в матрице D_B занулится $p - r$ строк, и тогда соотвествующее разложение мы можем написать по j от 1 до r и по i от r до n.

Если $\mathrm{rk}(A) = l < n$, то в матрице D_A занулится $n - l$ столбцов, разложение при этом мы сможем выписать только до l-ой координаты.

Обратим внимание, что построить $GSVD$-разложение - трудоемкая и дорогостоящая задача.

Метод взвешиванием

Проводят следующие действия с исходной задачей

$$\begin{cases} A\mathbf{x} = \mathbf{b} & \text{в смысле ЗНК} \\ B\mathbf{x} = \mathbf{d} & \text{как СЛАУ} \end{cases} \Leftrightarrow \begin{cases} A\mathbf{x} = \mathbf{b} \\ \lambda B\mathbf{x} = \lambda\mathbf{d} \end{cases} \to \begin{cases} A\mathbf{x} = \mathbf{b} & \text{ЗНК} \\ \lambda B\mathbf{x} = \lambda\mathbf{d} & \text{ЗНК} \end{cases} \Leftrightarrow \begin{pmatrix} A \\ \lambda B \end{pmatrix}\mathbf{x} = \begin{pmatrix} \mathbf{b} \\ \lambda\mathbf{d} \end{pmatrix} \text{ ЗНК}$$

То есть пытаются найти следующее

$$\inf_{\tilde{\mathbf{x}}} \left(\|A\tilde{\mathbf{x}} - b\|_2 + \|\lambda B\tilde{\mathbf{x}} - \lambda\mathbf{d}\|_2\right) \tag{1}$$

Утверждается, что при $\lambda \gg 1$ такой алгоритм сходится к решению исходной задачи.

Пусть имеется $GSVD$-разложение для этих матриц. Тогда выражение (1) можно переписать следующим образом

$$\inf_{\tilde{\mathbf{x}}} \left(\left\|UD_AX^{-1}\tilde{\mathbf{x}} - b\right\|_2^2 + \left\|\lambda VD_BX^{-1}\tilde{\mathbf{x}} - \lambda\mathbf{d}\right\|_2^2\right) \underset{\tilde{\mathbf{y}} = X^{-1}\tilde{\mathbf{x}}}{=} \inf_{\tilde{\mathbf{y}}} \left(\left\|D_A\tilde{\mathbf{y}} - U^T\mathbf{b}\right\|_2^2 + \left\|\lambda D_B\tilde{\mathbf{y}} - \lambda V^T\mathbf{d}\right\|_2^2\right)$$

Перепишем полученное равенство выражение покоординатно, пользуясь определением нормы

$$\sum_{i=1}^{n}(\alpha_i y_i - (\mathbf{u}_i, \mathbf{b}))^2 + \sum_{i=n+1}^{m}(\mathbf{u}_i, \mathbf{b})^2 + \lambda^2\sum_{i=1}^{p}(\beta_i y_i - (\mathbf{v}_i, \mathbf{b}))^2 \to \inf$$

Найдем y_i, на которых достигается inf. Продифференцируем $i = 1, \dots, p$

$$2\alpha_i(\alpha_i y_i - (\mathbf{u}_i, \mathbf{b})) + 2\lambda^2\beta_i(\beta_i y_i - (\mathbf{v}_i, \mathbf{b})) = 2\alpha_i^2 y_i - 2\alpha_i(\mathbf{u}_i, \mathbf{b}) + 2\lambda^2\beta_i^2 y_i - 2\lambda^2\beta_i(\mathbf{v}_i, \mathbf{b}) = 0$$

$$y_i(\alpha_i^2 + \lambda^2\beta_i^2) = \alpha_i(\mathbf{u}_i, \mathbf{b}) + \lambda^2\beta_i(\mathbf{v}_i, \mathbf{b}) \Rightarrow y_i = \frac{\alpha_i(\mathbf{u}_i, \mathbf{b}) + \lambda^2\beta_i(\mathbf{v}_i, \mathbf{b})}{\alpha_i^2 + \lambda^2\beta_i^2},\ i = 1, \dots, p$$

Продифференцируем $i = p+1, \dots, n$:

$$2\alpha_i(\alpha_i y_i - (\mathbf{u}_i, \mathbf{b})) = 0 \Rightarrow y_i = \frac{(\mathbf{u}_i, \mathbf{b})}{\alpha_i},\ i = p+1, \dots, n$$

Получаем итоговый ответ

$$\mathbf{x} = X\mathbf{y} = \sum_{j=1}^{p} \frac{\alpha_j(\mathbf{u}_j, \mathbf{b}) + \lambda^2\beta_j(\mathbf{v}_j, \mathbf{b})}{\alpha_j^2 + \lambda^2\beta_j^2}\mathbf{x}_j + \sum_{i=p+1}^{n} \frac{(\mathbf{u}_i, \mathbf{b})}{\alpha_i}\mathbf{x}_i$$

Отсюда подтверждается сходимость

$$\frac{\frac{\alpha_j(\mathbf{u}_j, \mathbf{b})}{\lambda^2} + \beta_j(\mathbf{v}_j, \mathbf{b})}{\frac{\alpha_j^2}{\lambda^2} + \beta_j^2}\mathbf{x}_j \underset{\lambda\to\infty}{\to} \frac{(\mathbf{v}_j, \mathbf{b})}{\beta_j}\mathbf{x}_j,\ j = 1, \dots, p$$

Замечание 30.3. При $\mathrm{rk}(B) = r < p$ в матрице D_B занулится $p - r$ строк, и тогда соотвествующее разложение мы можем написать по j от 1 до r и по i от r до n.

Если $\mathrm{rk}(A) = l < n$, то в матрице D_A занулится $n - l$ столбцов, разложение при этом мы сможем выписать только до l-ой координаты.