Прямые методы решения систем линейных алгебраических ур-ий.

$$A = \{a_{i,j}\}_{i,j=1}^{n}; \quad \vec{f} = (f_1, f_2, \ldots, f_n);$$

$$\vec{u} = (u_1, u_2, \ldots, u_n);$$

(1) $$A\vec{u} = \vec{f};$$

(1') $$\sum_{j=1}^{n} a_{i,j} u_j = f_i; \quad i = \overline{1,n};$$

Метод Гаусса. Прямой ход:

$$a_{i,1}\quad \begin{cases} u_1 + \sum\limits_{j=2}^{n} a^1_{1,j} u_j = f^1_1 ; \\ \sum\limits_{j=2}^{n} a^1_{i,j} u_j = f^1_i \quad (i=\overline{2,n}). \end{cases}$$

$$\Downarrow$$

$$\begin{cases} u_2 + \sum\limits_{j=3}^{n} a^2_{2,j} u_j = f^2_2 \\ \sum\limits_{j=3}^{n} a^2_{i,j} u_j = f^2_i \quad (i=\overline{3,n}) \end{cases}$$

$$(2)\quad \begin{cases} u_i + \sum\limits_{j=i+1}^{n} a^i_{i,j} u_j = f^i_i \quad (i=\overline{1,n-1}). \\ u_n = f^n_n \end{cases}$$

Обратный ход.

$$(3)\quad \begin{cases} u_n = f_n^n; \\ u_i = f_i^i - \sum\limits_{j=i+1}^{n} a_{i,j}^i u_j; \quad i = n-1, n-2, \dots, 1. \end{cases}$$

Теорема 1. Метод Гаусса (2), (3) может быть реализован

$\iff$ все главные миноры матрицы A отличны от 0 ⊕

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$$

Метод прогонки.

$$
(4)\quad \begin{cases} B_1 u_1 - C_1 u_2 = f_1 \\ -A_i u_{i-1} + B_i u_i - C_i u_{i+1} = f_i \quad (i = \overline{2, n-1}) \\ -A_n u_{n-1} + B_n u_n = f_n . \end{cases}
$$

$$
\begin{pmatrix} B_1 & -C_1 & \cdots & & 0 \\ -A_2 & B_2 & -C_2 & & \vdots \\ 0 & -A_3 & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & \\ 0 & \cdots & 0 & -A_n & B_n \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_n \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ \vdots \\ f_n \end{pmatrix}
$$

Метод Гаусса для системы (4) = <u>метод прогонки.</u>

$$u_1 = \frac{C_1}{B_1} u_2 + \frac{f_1}{B_1} \equiv \alpha_1 u_2 + \beta_1;$$

$$-A_2 \underline{(\alpha_1 u_2 + \beta_1)} + B_2 u_2 - C_2 u_3 = f_2$$

$$u_2 = \frac{C_2}{B_2 - A_2\alpha_1} \cdot u_3 + \frac{f_2 + \beta_1 A_2}{B_2 - A_2 \alpha_1} \equiv \alpha_2 u_3 + \beta_2;$$

Пусть $u_{i-1} = \alpha_{i-1} u_i + \beta_{i-1}$; Найдем "$u_i$".

$$-A_i \underline{(\alpha_{i-1} u_i + \beta_{i-1})} + B_i u_i - C_i u_{i+1} = f_i;$$

$$u_i = \frac{C_i}{B_i - A_i \alpha_{i-1}} \cdot u_{i+1} + \frac{f_i + \beta_{i-1} A_i}{B_i - A_i \alpha_{i-1}} \equiv \alpha_i u_{i+1} + \beta_i;$$

u_n найдем из системы ур-ий:

$$\begin{cases} u_{n-1} - \alpha_{n-1} u_n = \beta_{n-1}; \\ -A_n u_{n-1} + B_n u_n = f_n. \end{cases}$$

$$\Downarrow$$

$$\boxed{u_n = \frac{f_n + \beta_{n-1} A_n}{B_n - A_n \alpha_{n-1}}};$$

<u>Стандартная (монотонная) прогонка:</u>

I. $\alpha_1 = \dfrac{C_1}{B_1}; \ \beta_1 = \dfrac{f_1}{B_1};$

(5) $$\alpha_i = \frac{C_i}{B_i - A_i \alpha_{i-1}}; \quad \beta_i = \frac{f_i + \beta_{i-1} A_i}{B_i - A_i \alpha_{i-1}}; \quad i = \overline{2, n-1};$$

α_i, β_i — "прогоночные коэффициенты".

II. $$u_n = \frac{f_n + \beta_{n-1} A_n}{B_n - A_n \alpha_{n-1}};$$

(6)

$$u_i = \alpha_i u_{i+1} + \beta_i, \quad i = n-1, n-2, \dots, 1.$$

Определение 1. ① Прогонка (5), (6) наз. "корректной", если $B_1 \neq 0$; $B_i - A_i \alpha_{i-1} \neq 0$ $(i = \overline{2, n-1})$.

② Прогонка (5), (6) наз. "устойчивой", если:

$$|\alpha_i| \leq 1 \quad (i = \overline{1, n-1}) \quad (*)$$

Пусть α_i, β_i $(i = \overline{1, n-1})$ - вычисл. точно; $\overline{u}_n = u_n + \varepsilon_n$;

$\overset{(6)}{\Rightarrow} \overline{u}_i = u_i + \varepsilon_i$; $\overset{(6)}{\Rightarrow}$ $\varepsilon_i = \alpha_i \varepsilon_{i+1}$; $i = n-1, n-2, \dots$

$$|\varepsilon_i| \leq |\alpha_i| \cdot |\varepsilon_{i+1}| \leq |\varepsilon_{i+1}|.$$

Теорема 2. Рассмотрим систему ур-ний (4). Пусть:

1. $B_1 \neq 0,\ B_n \neq 0;\quad A_i \neq 0,\ C_i \neq 0 \quad (i=\overline{2, n-1})$.
2. $|B_1| \geqslant |C_1|,\quad |B_i| \geqslant |A_i| + |C_i| \quad (i=\overline{2, n-1});$

$$|B_n| \geqslant |A_n|.$$

и хотя бы одно из этих нер-в строгое.

$\Downarrow$

стандартная прогонка (5), (6) – корректна и устойчива.

Доказательство.

II. Предположим, что:

$$B_j - A_j \alpha_{j-1} \neq 0 \text{ и } |\alpha_j| \le 1 \text{ для } j = 1, 2, \dots i-1$$

при некотором фиксированном $i \le n-1$.

Докажем, что: $B_i - A_i \alpha_{i-1} \neq 0$ и $|\alpha_i| \le 1$.

$$|B_i - A_i \alpha_{i-1}| \ge |B_i| - |A_i| \cdot |\alpha_{i-1}| \overset{②}{\ge} |A_i| + |C_i| - |A_i| \cdot |\alpha_{i-1}|$$

$$= (1 - |\alpha_{i-1}|) \cdot |A_i| + |C_i| \ge |C_i| \overset{①}{>} 0.$$

$$\Downarrow \qquad \overset{\vee\!\!/}{0}$$

$$B_i - A_i \alpha_{i-1} \neq 0.$$

$$|\alpha_i| = \frac{|C_i|}{|B_i - A_i \alpha_{i-1}|} \le \frac{|C_i|}{|C_i|} = 1 \quad \oplus$$

Осталось док-ть: $B_n - A_n \alpha_{n-1} \neq 0$:

$$|B_n - A_n \alpha_{n-1}| \geq |B_n| - |A_n| \cdot \underset{\overset{\wedge}{1}}{|\alpha_{n-1}|} \geq |B_n| - |A_n| > 0.$$

$+$

Прогонка Дорра.

$$\begin{cases} \alpha_i = \dfrac{C_i}{B_i - A_i \alpha_{i-1}}; \quad \beta_i = \dfrac{f_i + A_i \beta_{i-1}}{B_i - A_i \alpha_{i-1}}; \\ \\ u_i = \alpha_i \cdot u_{i+1} + \beta_i; \end{cases}$$

$$\alpha_i \longrightarrow 1 - \alpha_i.$$

$$1 - \alpha_i = \frac{C_i}{B_i - A_i(1 - \alpha_{i-1})} \Rightarrow \alpha_i = 1 - \frac{C_i}{B_i - A_i + A_i \alpha_{i-1}} =$$

$$= \frac{B_i - A_i - C_i + A_i \alpha_{i-1}}{B_i - A_i + A_i \alpha_{i-1}} = \frac{S_i + A_i \alpha_{i-1}}{C_i + S_i + A_i \alpha_{i-1}};$$

где $S_1 \equiv B_1 - C_1$; $S_i \equiv B_i - A_i - C_i$ $(i = \overline{2, n-1})$;

$S_n \equiv B_n - A_n$.

Обычно выполнено: $S_i \geqslant 0$ $(i = \overline{1, n})$; $C_i > 0$; $A_i > 0$ $(i = \overline{2, n-1})$

В этом случае знаменатели в ф-лах прогонки Дорра содержат суммы положительных величин.

Формулы прогонки Дорра:

I. $$\alpha_1 = \frac{C_1}{B_1} = \frac{C_1}{S_1 + C_1}; \quad \beta_1 = \frac{f_1}{C_1 + S_1};$$

$$\alpha_i = \frac{S_i + A_i \alpha_{i-1}}{C_i + S_i + A_i \alpha_{i-1}}; \quad \beta_i = \frac{f_i + A_i \beta_{i-1}}{C_i + S_i + A_i \alpha_{i-1}}; \quad i = \overline{2, n-1}$$

II. $$u_n = \frac{f_n + A_n\alpha_{n-1}}{S_n + A_n\alpha_{n-1}};$$

$$u_i = (1-\alpha_i)\cdot u_{i+1} + \beta_i; \quad i = n-1, n-2, \dots, 1.$$ +

__Задача.__ Доказать, что число операций умножения и деления в прогонках "стандартная" и "Дорра" имеет порядок $O(n)$ при $n \to \infty$.

__Пример.__
$$\begin{cases} u_1 = \varphi, \\ -u_{i-1} + u_i - u_{i+1} = 0, \quad i = \overline{2, n-1}. \\ u_n = \psi. \end{cases}$$

Решение: $$u_i = \frac{\varphi\cdot\sin\frac{\pi(n-i)}{3} + \psi\cdot\sin\frac{\pi(i-1)}{3}}{\sin\frac{\pi(n-1)}{3}}; \quad n \neq 3k+1.$$

Вычисляем $\alpha_i = \dfrac{C_i}{B_i - A_i \alpha_{i-1}}$;

$$\alpha_1 = \frac{C_1}{B_1} = 1; \quad \alpha_2 = \frac{C_2}{B_2 - A_2 \cdot \alpha_1} = \frac{1}{1 - 1 \cdot 1} = \frac{1}{0} \; (?)$$

$$B_i - A_i - C_i = 1 - 1 - 1 = -1 < 0; \quad i = \overline{2, n-1};$$

$\uparrow$ условия (2) диагонального преобладания не выполнены.

Погрешности округления.

(1) $\boxed{q = (-1)^s \cdot M \cdot r^p;}$

$r \in \mathbb{Z} \cap (1, +\infty)$; основание системы счисления;

$p \in \mathbb{Z} \cap (-p_*, p^*)$ – порядок числа "q"; p_*, p^* – целые и положительные (границы порядка)

$M = \sum_{i=1}^{t} q_i r^{-i}$ – мантисса числа "q";

$q_i \in \mathbb{Z} \cap [0, r-1]$ – разряды мантиссы; $i = \overline{1,t}$;

t – целое и положительное (разрядность)

$q_1 \neq 0$ – для нормализованного представления.

$S = 0 \vee 1$ – знак числа q

Q^+ – положит. компонента мн-ва (1).

q_{min} и q_{max} – во мн-ве Q^+:

$$q = r^p \cdot \left(\sum_{i=1}^{t} q_i r^{-i}\right)$$

$$-p_* < p < p^* ; \quad 0 \le q_i \le r-1 ;$$

$$q_{min}: \quad p = -p_* ; \quad q_1 = 1 ; \quad q_i = 0 \ (i = \overline{2,t})$$

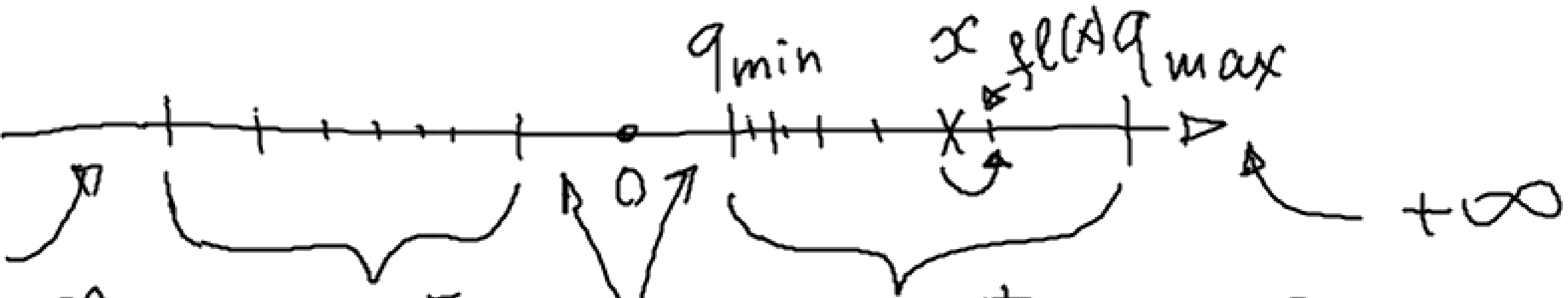
$$q_{min} = r^{-p_*} \cdot r^{-1} = \underline{\underline{r^{-(p_*+1)}}}$$

$$\boxed{q_{max} = r^{p^*}}$$

$$q_{max}: \quad p = p^* ; \quad q_i = r-1 ; \quad i = \overline{1,t} ;$$

$$q_{max} = r^{p^*} \cdot \left(\sum_{i=1}^{t} r^{-i}\right)(r-1) = \underline{\underline{r^{p^*} \cdot (1 - r^{-t})}}$$

$$\sum_{i=1}^{t} \left(\frac{1}{r}\right)^i = \frac{1 - \left(\frac{1}{r}\right)^{t+1}}{1 - 1/r} - 1 = \frac{1/r - (1/r)^{t+1}}{1 - 1/r} = \frac{1 - r^{-t}}{r - 1} ;$$



$Q^- \quad 0 \quad Q^+ \subseteq Q$ – рацион. числа

$x \in \mathbb{R} \Rightarrow$ (1) $|x| \geq q_{max} \Rightarrow x = +\infty$.

(2) $|x| \leq q_{min} \Rightarrow x = 0$;

(3) $q_{min} < |x| < q_{max} \Rightarrow$

$\Rightarrow x \longrightarrow fl(x) \in Q^- \cup Q^+$

$\dfrac{fl(x) - x}{x}$ – относительная погрешность округления.

Теорема 1. $q_{min} < |x| < q_{max} \Longrightarrow \left| \frac{fl(x) - x}{x} \right| \le \frac{r^{1-t}}{2}$;

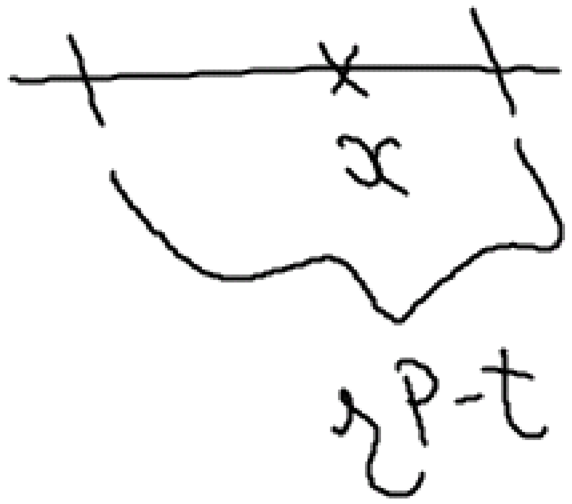
Док-во. $q_{min} < x < q_{max}$ – считаем, без огр. общности.

$$q = r^p \left(\sum_{i=1}^{t} q_i r^{-i} \right) \in [r^{p-1}, r^p];$$

$x \in [r^{p-1}, r^p]$ с некоторым "p"

Т.к. $[q_{min}, q_{max}] = \bigcup_{p=-p_*}^{p^*} [r^{p-1}, r^p]$;

$$\begin{cases} |fl(x) - x| \le \frac{r^{p-t}}{2}; \\ |x| \ge r^{p-1}. \end{cases}$$



$$\Downarrow$$

$$\frac{|fl(x) - x|}{|x|} \le \frac{r^{p-t}}{2 r^{p-1}} = \frac{r^{1-t}}{2} \quad (+)$$

Форматы представления вещ. чисел в памяти комп.

$$q = (-1)^s \left(\sum_{i=1}^{t} q_i 2^{-i} \right) 2^p ; \quad p = \sum_{i=0}^{k-1} p_i 2^i - p_* ;$$

$$1 + k + t - 1 = k + t$$

$$\underbrace{\boxed{s}}\underbrace{\boxed{p_0}\boxed{p_1}\boxed{\;\dots\;}\boxed{p_{k-1}}}_{P}\underbrace{\boxed{q_2 \dots q_i}\boxed{\;\dots\;}\boxed{q_t}}_{M}$$

(E1): $k = 8$, $t = 24$, $p_* = 126$; (одинарная точность)

(E2) $k = 11$, $t = 53$, $p_* = 1022$; (двойная точность)

$$p^* = 2^k - p_* - 1 ;$$

Для (E1): $q_{min} = 2^{-(p_*+1)} = 2^{-127} \approx 10^{-38}$; — маш. ноль машин. бесконеч.: $q_{max} = 2^{p^*} = 2^{129} \approx 10^{39}$;

машин. "ε": $\varepsilon = 2^{-t} = 2^{-24} \approx 10^{-7}$;

Для (E2): $q_{min} = 2^{-1023} \approx 10^{-308}$; $(\emptyset)$

$q_{max} = 2^{1025} \approx 10^{309}$; (∞)

$\varepsilon = 2^{-53} \approx 10^{-16}$;

$x, y \in \mathbb{R} \Rightarrow fl(x), fl(y) \in Q^- \cup Q^+ \Rightarrow$

$\Rightarrow fl(x) \oplus fl(y) \Rightarrow fl[fl(x) \oplus fl(y)]$;

$p_1 + p_2 \geq p^*$ – "переполнение";

$-p_1 - p_2 \leq -(p_* + 1)$ – "потеря порядка";

$\oplus$ – операция "+" или "–".

$x, y \in \mathbb{R}$;
$$\frac{fl(x) - x}{x} = \varepsilon_x \Rightarrow \underline{fl(x) = (\varepsilon_x + 1) \cdot x};$$

$$\frac{fl(y) - y}{y} = \varepsilon_y \Rightarrow \underline{fl(y) = (\varepsilon_y + 1) \cdot y};$$

$$\frac{fl\left[fl(x) \oplus fl(y)\right] - fl(x) \oplus fl(y)}{fl(x) \oplus fl(y)} = \varepsilon$$

$$\underline{fl\left[fl(x) \oplus fl(y)\right] = (\varepsilon + 1) \cdot \left[fl(x) \oplus fl(y)\right]};$$

$$\frac{fl\left[fl(x) \oplus fl(y)\right] - (x \oplus y)}{x \oplus y} = \frac{(\varepsilon + 1)\left[(\varepsilon_x + 1)x \oplus (\varepsilon_y + 1)y\right] - x \oplus y}{x \oplus y} =$$

$$= \varepsilon + (\varepsilon+1)\,\frac{\varepsilon_x \cdot x \oplus \varepsilon_y \cdot y}{x \oplus y} \equiv E;$$

$\underline{\underline{\oplus = +}}$: $\quad \varepsilon, \varepsilon_x, \varepsilon_y \approx \dfrac{r^{1-t}}{2}$;

$$|E| \leq \frac{r^{1-t}}{2} + \left(1 + \frac{r^{1-t}}{2}\right)\cdot\frac{r^{1-t}}{2} \approx \frac{r^{1-t}}{2} \quad (t \to \infty)$$

$\underline{\underline{\oplus = -}}$; $\quad \dfrac{\varepsilon_x \cdot x - \varepsilon_y \cdot y}{x-y} \approx E$

проблема : $x-y$ – малое.

<u>Пример.</u> $z^2 - 200z + 1 = 0$ – найти наименьший корень.

$\Rightarrow \quad z = 100 - \sqrt{9999}$; $\quad \boxed{r = 10 \text{ и } t = 4;}$

$$\begin{cases} x = 100; \\ y = \sqrt{9999} \approx \underbrace{99,99}499987\ldots; \\ x - y = 0,00500012\ldots \end{cases}$$

$$\begin{cases} fl(x) = 100; \\ fl(y) = 99,99; \\ fl(x) - fl(y) = 0,01; \\ fl[fl(x) - fl(y)] = 0,01; \end{cases}$$

$$E = \frac{0,01 - 0,00500012\ldots}{0,00500012\ldots} = 0,99992499\ldots \approx 1;$$

$$\varepsilon_x = 0; \; \varepsilon_y = 5 \cdot 10^{-5}$$

$$x - y = 5 \cdot 10^{-3}; \; y \approx 100$$

$$z = \frac{1}{100+\sqrt{9999}};$$

$$fl(x) + fl(y) = 199,99;$$

$$fl[fl(x)+fl(y)] = 200;$$

$$\boxed{z = 0,005;}$$