Курс «Численные методы»

ВОЛКОВ Василий Михайлович, Минск, БГУ
v.volkov@tut.by

Минск, 4 сентября 2019

Зачем изучать численные методы?

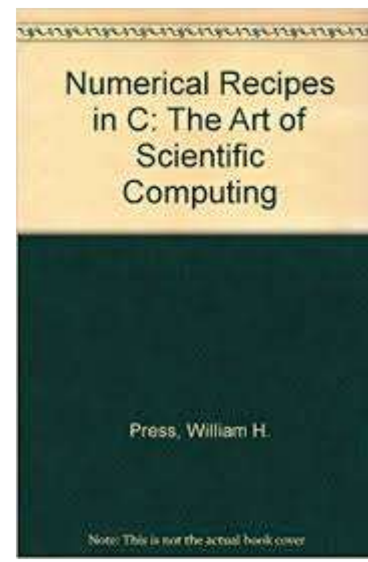
«Потому, что все оттенки смысла
Умное число предает…» Н.Гумилев 1921 г.

XXI век – век цифровых технологий

…

…

Численные методы – это цифровые технологии в математике, как языке современной науки.



Floating point operations per second (FLOPS). TOP 500

2016. Тяньхэ-2: 33.86(54.9) пентафлопс (10^{15} flops)
2017 TaihuLight 93 petaflops

1.0 Компьютерная арифметика Вычислительная погрешность

1.0.1 Формат чисел с плавающей запятой (floating point)

$$a = \pm d^{s}\underbrace{\left(\alpha_1 d^{-1} + \alpha_2 d^{-2} + \cdots + \alpha_m d^{-m}\right)}_{\text{МАНТИССА}},$$

$$\alpha_1 \neq 0, \quad 0 \leq \alpha_k < d$$

Например: d=10 (десятичная система), **0.31415e+001**

Длина мантиссы определяет относительную погрешность представления чисел: «**машинное епсилон -- ε**» минимальное число, которое, будучи прибавленным к единице, делает сумму отличной от единицы. Другими словами, это относительная погрешность представления чисел с плавающей запятой (точкой)

1.0.2 Форматы single и double. Стандарт IEEE 756

	длина	ε	∞
single	32 бит	2^{-23} =1.19e-7	>3.4028e+38
double	64 бит	2^{-52} =2.22e-16	>1.7977e+308

Особенности компьютерных вычислений (double)

Точная арифметика	Компьютерная арифметика
2+ε-2 = ε	2+ε-2 = 0
$\sum_{n=1}^{\infty} \frac{1}{n} = \infty$	$\sum_{n=1}^{\infty} \frac{1}{n} < 49$
$10^{-10} \sum_{n=1}^{10^{10}} 0.2 - 0.2 = 0$	$10^{-10} \sum_{n=1}^{10^{10}} 0.2 - 0.2 = 3.2625e - 08$
$\sqrt{(10^{160})^2} = 10^{160}$	$\sqrt{(10^{160})^2} = Inf$

Вопросы

Что получится при вычислениях в классе double?:

$$\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n, \; n = 10^{20}.$$

Лекция 1.

1.1 Нормы векторов и матриц. Оценка погрешности решения систем ЛАУ. Число обусловленности

Аксиомы нормы

Вектор $\mathrm{x} = (x_1, x_2, x_3, \ \ldots, x_N)^T, \quad \mathbf{x} \in R^N$

Норма вектора: $\|\cdot\| = R^N \to R$:

1. $\|\mathbf{x}\| \geq 0, \quad \|\mathbf{x}\| = 0 \Leftrightarrow \mathbf{x} = 0$ – *положительная определенность*;
2. $\|\alpha \mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\|, \quad \forall \alpha \in R$ – *линейность при умножении на скаляр*;
3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ – *неравенство треугольника*.

Примеры:

$$\|x\|_2 = \sqrt{\sum_{m=1}^{N} x_m{}^2}$$ – Евклидова норма

$$\|x\|_\infty = \max_m |x_m|$$ – максимальная норма

$$1/\sqrt{N}\,\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2$$ – Эквивалентность норм

Матрицы и нормы матрицы

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1M} \\ a_{21} & a_{22} & \cdots & a_{2M} \\ \cdots & \cdots & a_{km} & \cdots \\ a_{K1} & a_{K2} & \cdots & a_{KM} \end{bmatrix}. \qquad \boldsymbol{A} \in \boldsymbol{R}^{N \times N}.$$

Норма матрицы A, подчиненная векторной норме $\|\cdot\|$, определяется числом

$$\| A \| = \sup_{\mathbf{x} \neq 0} \frac{\| A\mathbf{x} \|}{\| \mathbf{x} \|} = \sup_{\|\mathbf{x}\|=1} \| A\mathbf{x} \|.$$

Подчиненные нормы обладают свойством мультипликативности:

$$\| Ax \| \leq \| A \| \cdot \| \mathbf{x} \|$$

Примеры матричных норм

$$\| A \|_{\infty} = \max_{k} \left\{ \sum_{m=1}^{N} | a_{km} | \right\}$$

$$\| A \|_{2} = \max\{\lambda(AA^{T})^{1/2}\}$$

$\max\{\lambda(AA^{T})$ Максимальное сингулярное число

Оценка погрешности решения систем ЛАУ

Система линейных алгебраических уравнений в матричном виде

$$\mathbf{Ax} = \mathbf{f}. \quad \sum_{m=1}^{N} a_{km} x_m = f_k, \; k = \overline{1, N}.$$

$$\mathbf{A} \in \mathbf{R}^{\mathbf{N}\times\mathbf{N}}; \quad \mathbf{x}, \mathbf{f} \in \mathbf{R}^{\mathbf{N}},$$

Пусть найдено приближенное решение системы

$$\tilde{\mathbf{x}} = \mathbf{x} + \delta\mathbf{x}$$

Приближенное решение исходной системы является точным решением некоторой ***возмущенной*** системы

$$A(\mathbf{x} + \delta\mathbf{x}) = \mathbf{f} + \delta\mathbf{f}$$

$$\boxed{\delta\mathbf{f} = A(\mathbf{x} + \delta\mathbf{x}) - \mathbf{f}}$$ – **невязка.**

Оценка погрешности возмущенной задачи

$$A\delta\mathbf{x} = \delta\mathbf{f} \quad \Rightarrow \quad \|\delta\mathbf{x}\| = \|A^{-1}\delta\mathbf{f}\| \quad \Rightarrow \quad \|\delta\mathbf{x}\| \le \|A^{-1}\| \cdot \|\delta\mathbf{f}\|$$

Проделывая то же самое с исходной системой будем иметь

$$A\mathbf{x} = \mathbf{f} \Rightarrow \quad \|A\| \cdot \|\mathbf{x}\| \ge \|\mathbf{f}\| \quad \Rightarrow \quad \|\mathbf{x}\| \ge \|A\|^{-1} \cdot \|\mathbf{f}\|$$

Деление первого неравенства на второе дает оценку

$$\boxed{\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \le \|A^{-1}\| \cdot \|A\| \cdot \frac{\|\delta\mathbf{f}\|}{\|\mathbf{f}\|}}$$

<u>Число обусловленности матрицы</u>

$$\boxed{M_A = \mathrm{cond}(A) = \|A\| \cdot \|A^{-1}\|}$$

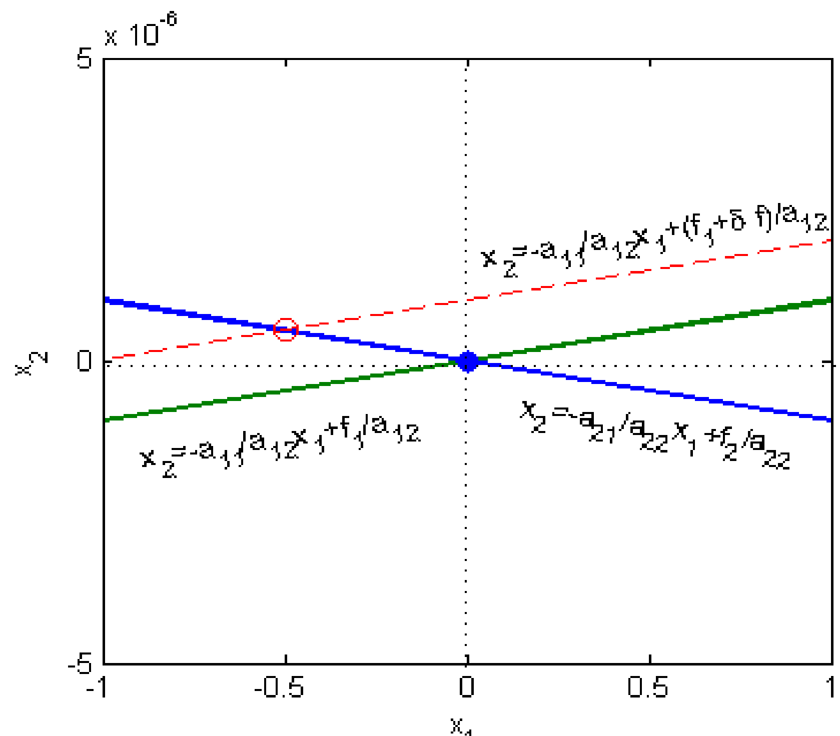
Геометрический смысл плохой обусловленности

$$\begin{cases} a_{11}x_1 + a_{12}x_2 = f_1, \\ a_{21}x_1 + a_{22}x_2 = f_2. \end{cases} \qquad \tan(\alpha_1) = -\frac{a_{11}}{a_{12}}, \qquad \tan(\alpha_2) = -\frac{a_{21}}{a_{22}}$$

Определитель матрицы $\det(A) = a_{11}a_{22} - a_{12}a_{21}.$

$$\boxed{\det(A) = a_{11}a_{22} - a_{12}a_{21} = 0 \Leftrightarrow \tan(\alpha_1) = \tan(\alpha_2)}$$

$$a_{11} = -a_{21} = -10^{-6},$$
$$a_{22} = a_{12} = 0,$$
$$f_1 = f_2 = 1$$

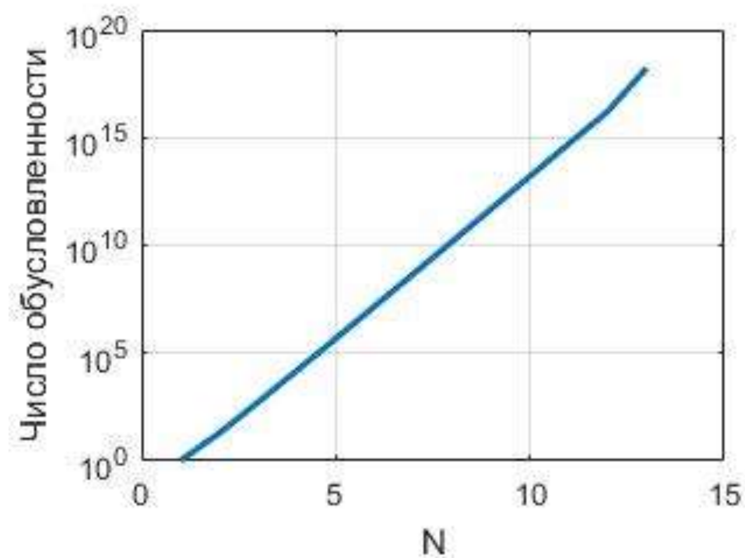


Пример плохо обусловленной матрицы

Матрица Гильберта

$$H_n = \begin{pmatrix} 1 & 1/2 & 1/3 & \cdots & 1/n \\ 1/2 & 1/3 & 1/4 & \cdots & 1/n+1 \\ 1/3 & 1/4 & 1/5 & \cdots & 1/n+2 \\ \vdots & \vdots & \vdots & & \vdots \\ 1/n & 1/n+1 & 1/n+2 & \cdots & 1/2n-1 \end{pmatrix}$$

Зависимость числа обусловленности от размерности матрицы Гильберта



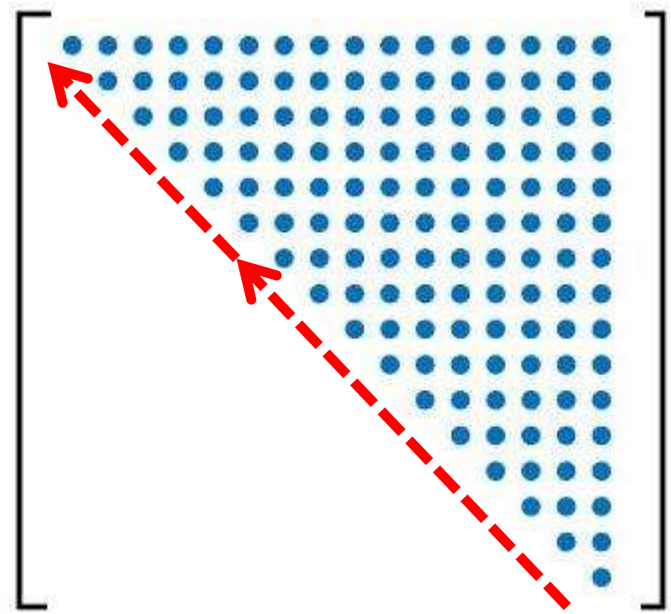
Решение системы ЛАУ с матрицей Гильберта 14x14 имеем предупреждение :

Warning: Matrix is close to singular or badly scaled. Results may be inaccurate. RCOND = 1.602620e-18.

Метод Гаусса

Предпосылкой для метода Гаусса является тот факт, что для системы с треугольной матрицей решение находится относительно просто.

$$a_{k\le m} \ne 0, \quad a_{k>m} = 0$$



$$\sum_{m=k}^{N} a_{km} x_m = f_k$$

$$x_N = f_N / a_{NN}$$

Обратный ход метода Гаусса

$$x_k = \left(f_k - \sum_{m=k+1}^{N} a_{km} x_m \right) a_{kk}^{-1}, \qquad k = N-1, N-2, \ldots, 2, 1$$

Приведение матрицы к треугольному виду

Приведение матрицы системы к треугольному виду основано на **элементарных преобразованиях**:

- *Умножение строки на ненулевое число (масштабирование строки);*
- *Перестановка строк;*
- *Замена строки на ее линейную комбинацию с другими строками.*

Каждое из элементарных преобразований можно выразить в виде умножения левой и правой части системы на соответствующую (квадратную не вырожденную) матрицу преобразований. Матрицы элементарных преобразований имеют ту же размерность , что и матрица системы.

Масштабирование и перестановка строк.

<u>Элементарная матрица масштабирования</u> к-й строки получается из единичной матрицы, в которой единичному элементу $a_{kk} = 1$ присваивается значение масштабирующего множителя $a_{kk} = s$
Произведение произвольного числа элементарных матриц масштабирования тоже является матрицей масштабирования.

Элементарная матрица перестановок к-й и m-й строк получается и единичной матрицы путем перестановки в ней соответствующих (к-й и m-й) строк.

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 5 & 9 & 13 \\ 2 & 6 & 10 & 14 \\ 3 & 7 & 11 & 15 \\ 4 & 8 & 12 & 16 \end{bmatrix} = \begin{bmatrix} 1 & 5 & 9 & 13 \\ 4 & 12 & 20 & 2 \\ 3 & 7 & 11 & 15 \\ 4 & 8 & 12 & 16 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 1 & 5 & 9 & 13 \\ 2 & 6 & 10 & 14 \\ 3 & 7 & 11 & 15 \\ 4 & 8 & 12 & 16 \end{bmatrix} = \begin{bmatrix} 1 & 5 & 9 & 13 \\ 4 & 8 & 12 & 16 \\ 3 & 7 & 11 & 15 \\ 2 & 6 & 10 & 14 \end{bmatrix}$$

Элементарная треугольная матрица

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \cdots & \cdots & a_{km} & \cdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{bmatrix}.$$

$$L_{(k)} = \begin{bmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \ddots & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1/a_{kk} & 0 & \cdots & 0 \\ 0 & 0 & -a_{k+1,k}/a_{kk} & 1 & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \ddots & 0 \\ 0 & \cdots & -a_{N,k}/a_{kk} & \cdots & 0 & 1 \end{bmatrix}.$$

При умножении на матрицу $L_{(k)}$ элементы к-го столбца ниже диагонали обращаются в нуль, а элементы левее этого столбца остаются неизменными.

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ -3 & 0 & 1 & 0 \\ -4 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 5 & 9 & 13 \\ 2 & 6 & 10 & 14 \\ 3 & 7 & 11 & 15 \\ 4 & 8 & 12 & 16 \end{bmatrix} = \begin{bmatrix} 1 & 5 & 9 & 13 \\ 0 & -4 & -8 & -12 \\ 0 & -8 & -16 & -24 \\ 0 & -12 & -24 & -36 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1/4 & 0 & 0 \\ 0 & -2 & 1 & 0 \\ 0 & -3 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 5 & 9 & 13 \\ 0 & -4 & -8 & -12 \\ 0 & -8 & -16 & -24 \\ 0 & -12 & -24 & -36 \end{bmatrix} = \begin{bmatrix} 1 & 5 & 9 & 13 \\ 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Приведение матрицы к треугольному виду (Прямой ход метода Гаусса)

$$A_{(1)} = L_{(1)} A,$$

$$A_{(2)} = L_{(2)} A_{(1)},$$

....

$$A_{(N)} = L_{(N)} A_{(N-1)}$$

Приведение матрицы к треугольному виду сводится к N матричным умножениям. Каждое такое умножение требует порядка ***O(N²)*** арифметических операций. Общая вычислительная сложность алгоритма ***O(N³)*** . Для обратного хода метода Гаусса несложно показать, что его вычислительная сложность ***O(N²),*** что на порядок ниже, чем у прямого хода.

Выбор главного элемента

Элемент a_{kk}, определяющий элемент $l_{kk} = 1/a_{kk}$ элементарной треугольной матрицы $L_{(k)}$, принято называть ***главным***. Очевидно, что если $a_{kk} = 0$, то метод Гаусса становится некорректным (деление на ноль). Избежать данной ситуации можно с помощью перестановки строк таким образом, чтобы $|a_{m>k,k}| \le |a_{k,k}|$. Выбор главного элемента не только позволяет избежать деления на ноль, но и предотвращает деление на малое число, близкое к нулю, что препятствует катастрофическому возрастанию абсолютной погрешности. Если оказывается, что главный элемент и все элементы в столбце ниже главного равны нулю,

$$\max_m \{a_{m\le k,k}\} = 0.$$

то это говорит о том, что преобразуемая матрица ***вырожденная.***

СПАСИБО ЗА ВНИМАНИЕ!