Лекция по курсу «Машинная арифметика в рациональных числах» на тему интервальные вычисления

2020

Пример задачи потери точности

Производная функции f в точке x определяется формулой

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h},$$

поэтому при «малом» h

$$f'(x) \approx \frac{f(x+h) - f(x)}{h} \equiv \Delta_h f(x).$$

$$\Delta_h f(x) \equiv \frac{f(x+h) - f(x)}{h} = \frac{f(x) + hf'(x) + h^2 f''(\xi)/2 - f(x)}{h} =$$
$$= f'(x) + hf''(\xi)/2, \qquad x < \xi < x + h.$$

Таким образом, неустранимая ошибка этого приближения равна $hf''(\xi)/2$.

В численном анализе ошибка нередко зависит от некоторого параметра, каким в данном примере было h. Во многих случаях достаточно указать характер зависимости ошибки от параметра, не выписывая выражение для ошибки более детально. В подобных случаях используется введенное Бахманом и Ландау обозначение $O(h)$, так называемое «O большое»; оно означает лишь, что неустранимая ошибка стремится к нулю не медленнее, чем само h. Более точно, мы будем писать

$$p(t) = O(q(t)) \quad \text{при} \quad t \to t_0,$$

Пример задачи потери точности

Нужно вычислить все четыре корня полинома

$$x^4 - 4x^3 + 8x^2 - 16x + 15.99999999 = (x - 2)^4 - 10^{-8} = 0.$$

Корни удовлетворяют уравнению $(x - 2)^2 = \pm 10^{-4}$ и, следовательно,

$$x - 2 = \pm\sqrt{\pm 10^{-4}} = \pm 10^{-2} \quad \text{или} \quad x - 2 = \pm 10^{-2} i.$$

Итак, корнями будут числа $x_1 = 2.01$, $x_2 = 1.99$, $x_3 = 2 + .01i$, $x_4 = 2 - .01i$. Если мы работаем на компьютере, для которого $\varepsilon_{маш} > 10^{-10}$, то свободный член полинома будет округлен до 16.0. С точки зрения компьютера, теперь решается уравнение

$$(x - 2)^4 = 0.$$

Решение системы линейных уравнений

$AX = B$

$X = A^{-1}B$

- Есть методы, которые дешевле
- дают более точные ответы
- Лучше сохраняют структуру матрицы
- Более инфомативны

Типы матриц

1. *Хранимая матрица*, т.е. матрица, все n^2 элементов которой хранятся в оперативной памяти машины. Это ограничивает значение порядка n несколькими сотнями для машин средней мощности и примерно тысячей на больших машинах.

2. *Разреженная матрица*, т.е. матрица, большинство элементов которой – нули, а ненулевые элементы могут или храниться в какой-либо специальной структуре данных, или регенерироваться по мере необходимости. Матрицы этого типа часто возникают в конечно-разностных и конечно-элементных методах решения дифференциальных уравнений с частными производными. Порядок n зачастую достигает нескольких десятков тысяч, а иногда и того больше. Приведем пример разреженной матрицы, элементы которой легко восстанавливаются:

$$\begin{pmatrix} 4 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 4 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 4 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 4 \end{pmatrix}$$

Невязка

Пусть x^* – вычисленное решение линейной системы $Ax = b$. Существуют две общеупотребительные меры погрешности в x^*: *вектор ошибки*

$$e = x - x^*$$

и *невязка*

$$r = b - Ax^* = A(x - x^*) = Ae.$$

Az = 0
Ax = b

A(x+αb) = 0

Обусловленность

Однако даже если невязка мала, это, вообще говоря, *не означает*, что мал и вектор ошибки. Связь между величиной невязки и величиной ошибки отчасти определяется характеристикой, называемой *числом обусловленности* матрицы A и обозначаемой $\operatorname{cond}(A)$. Действует приближенное равенство

$$\text{величина ошибки в решении} \approx \text{величина решения} \times \operatorname{cond}(A) \times \varepsilon_{\text{маш}}.$$

Обусловленность – это внутреннее свойство матрицы, не связанное с тем, как именно решается система уравнений. Матрицы с большими числами обусловленности дают большие ошибки при решении систем. Логарифм числа обусловленности допускает полезную интерпретацию: он приближенно равен числу значащих цифр, теряемых в решении системы $Ax = b$. Так, если $\operatorname{cond}(A) = 10^5$, а машинный эпсилон равен 10^{-8}, то лучшее, на что можно рассчитывать, – это приблизительно три верных разряда в компонентах вычисленного решения.

Плохая масштабируемость вектора

$$x = (10^5,\ 10^{-6})^T.$$

Если x – приближенное решение системы линейных уравнений, то его погрешность обычно пропорциональна наибольшей компоненте вектора, в данном случае числу 10^5. Если значение x_1 имеет четыре верных десятичных разряда, то ошибка в компоненте x_2 равна приблизительно $10^5 \times 10^{-4} = 10^1$, и неразумно рассчитывать на то, что x_2 имеет хоть один верный знак. Подобную ситуацию иногда удается исправить в момент формирования системы путем изменения единиц измерения.

Нормы векторов

Норма вектора x – это число, измеряющее общую величину элементов вектора. Она обозначается символом $\|x\|$. Задать норму можно многими способами. Теоретически всякая функция, удовлетворяющая приводимым ниже четырем условиям, приемлема в качестве нормы. Большинство теорем, в формулировках которых участвуют нормы, верны независимо от того, какие именно это нормы. Поэтому их формулировки можно читать, заменяя для большего удобства символ нормы символом длины вектора. Однако нормы используются также при вычислениях, и здесь выбор конкретной нормы имеет важное значение. Наиболее употребительная векторная норма – это евклидова длина, или 2-норма

$$\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2}.$$

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \qquad \|x\|_\infty = \max_i |x_i|.$$

Нормы векторов

$$M = \max_x \frac{\|Ax\|}{\|x\|} \quad (\Rightarrow \|Ax\| \leqslant M\|x\|),$$

$$m = \min_x \frac{\|Ax\|}{\|x\|} \quad (\Rightarrow m\|x\| \leqslant \|Ax\|).$$

Максимум и минимум берутся по всем ненулевым векторам. Заметим, что если матрица A вырождена, то $m = 0$. Отношение M/m называется *числом обусловленности* матрицы A:

$$\operatorname{cond}(A) = \frac{\max_x \dfrac{\|Ax\|}{\|x\|}}{\min_x \dfrac{\|Ax\|}{\|x\|}}.$$

Нормы векторов

Рассмотрим систему уравнений

$$Ax = b$$

и другую систему, полученную изменением правой части:

$$A(x + \Delta x) = b + \Delta b.$$

Будем считать Δb ошибкой в b, а Δx — соответствующей ошибкой в x, хотя нет необходимости предполагать, что ошибки малы. Поскольку $A(\Delta x) = \Delta b$, определения M и m немедленно ведут к неравенствам

$$\|Ax\| = \|b\| \leqslant M\|x\|$$

и

$$\|A\Delta x\| = \|\Delta b\| \geqslant m\|\Delta x\| .$$

Следовательно, при $M \neq 0$

$$\frac{\|\Delta x\|}{\|x\|} \leqslant \operatorname{cond}(A)\,\frac{\|\Delta b\|}{\|b\|} .$$

Нормы векторов

Величина $\|\Delta b\| / \|b\|$ есть *относительное* изменение правой части, а величина $\|\Delta x\| / \|x\|$ – *относительная* ошибка, вызванная этим изменением. Использование относительных изменений имеет то преимущество, что они безразмерны, т. е. нечувствительны к общим масштабирующим множителям.

Полученное неравенство показывает, что число обусловленности выполняет роль коэффициента увеличения относительной ошибки. Изменения правой части могут повлечь за собой изменения в решении, бóльшие в cond(A) раз. Оказывается, что то же самое справедливо в отношении изменений в коэффициентах матрицы.

$$\frac{\|b - A x^*\|}{\|A\| \, \|x^*\|} \leqslant C \varepsilon_{\text{маш}} ,$$

$$\frac{\|x - x^*\|}{\|x^*\|} \leqslant C \operatorname{cond}(A) \varepsilon_{\text{маш}}$$

Теорема Брауэра о неподвижной точке

Определение 1.7. *Элемент a множества A называют неподвижной точкой отображения $f\colon A \to A$, если $f(a) = a$.*

Более подробно, рассмотрим замкнутый шар в n-мерном пространстве $B^n \subset \mathbb{R}^n$. Пусть $f\colon B^n \to B^n$ — некоторое непрерывное отображение этого шара в себя (не обязательно строго внутрь себя, не обязательно биективное, т.е. даже не обязательно сюръективное). Тогда найдется такая точка $x \in B^n$, что $f(x) = x$.