Численные методы
Лекция 1

Евгений Александрович Яревский

28 октября 2020

evgeny.yarevsky@yandex.ru

Основная цель численного анализа и научных вычислений – разработка эффективных и точных методов вычисления величин, которые трудно или невозможно получить аналитически.

Когда нужны численные методы:

- Высокая точность результатов
- Большое количество/высокая скорость вычислений
- “Плохо обусловленные” модели явлений/процессов

Основная цель численного анализа и научных вычислений – разработка эффективных и точных методов вычисления величин, которые трудно или невозможно получить аналитически.

Когда нужны численные методы:

- Высокая точность результатов
- Большое количество/высокая скорость вычислений
- “Плохо обусловленные” модели явлений/процессов

$$x^4 - 2^4 = \varepsilon, \qquad x \approx (\pm 1)(\pm i)\, 2(1 + \varepsilon/64),$$
$$(x-2)^4 = \varepsilon, \qquad x = 2 + (\pm 1)(\pm i)\sqrt[4]{\varepsilon},$$

План курса

- Точность и погрешность.
- Решение систем линейных алгебраических уравнений.
- Решение задачи на собственные значения.
- SVD и CUR разложения матриц.

Источники погрешностей (ошибок)

Численные методы подвержены влиянию **погрешностей** (ошибок).

A) ошибки входных (исходных) данных – *систематические* и *случайные* ошибки

B) ошибки округления (представления чисел и операций в компьютере)

C) ошибки, возникающие из-за “обрезания” бесконечно малых или больших величин

D) упрощения в используемых математических моделях

E) “человеческие” и машинные ошибки

Ошибки типов A и D, как правило, не могут контролироваться в рамках численных методов.

Тип C, как правило, можно контролировать.

Тип B можно контролировать частично.

Понятия абсолютной ($A \pm \Delta A$) и относительной ($\Delta A/A$) погрешностей.
(Масса нейтрино: $m_\nu^2 = -22 \pm 17_{stat} \pm 17_{syst}$ эВ^2, 1995.
Сейчас: не более 0.28 эВ, но не 0 – Нобелевская премия по физике 2015).

Часто хочется иметь границы (точные или приближенные) для абсолютной и относительной погрешностей.
Получить их обычно трудно, так что часто пользуются оценками погрешностей.

Числа с плавающей точкой (запятой) (FP)

Дробные числа с плавающей точкой – удобная модель вещественных чисел: $\pm m \times \beta^E$.

Исключение неоднозначности представления – нормализованные числа

В компьютере занимают фиксированное число бит (байт) – конечное количество чисел!

Особенности FP по сравнению с вещественными числами:

- Если a, b принадлежат FP, то $a + b$, $a - b$, $a * b$, a/b не обязательно принадлежат FP
- Из $a + b = a$ не следует, что $b = 0$
- Нет ассоциативности: $a + (b + c) \neq (a + b) + c$
- Нет дистрибутивности: $a * (b + c) \neq a * b + a * c$

Коммутативность присутствует.

Пример.

Вычисления с одинарной точностью:

$1 + 1/2 + 1/3 + \ldots + 1/10^9 = 16.1$

$1/10^9 + 1/(10^9 - 1) + \ldots + 1/2 + 1 = 23.02$

Нормализованное представление вещественного числа a в форме с плавающей точкой:

$$a = \pm m \cdot \beta^e, \quad \beta^{-1} \leq m < 1, \quad e - \text{целое}.$$

m – мантисса, e – экспонента, β – основание системы.
В реальности, количество цифр в e и m ограничено.
Таким образом,

$$\overline{a} = \pm \overline{m} \cdot \beta^e, \quad \overline{m} = (0.d_1 d_2 \cdots d_t)_\beta, \quad 0 \leq d_i < \beta.$$

$$e_{min} \leq e \leq e_{max}.$$

Машинная точность ϵ_M – минимальное положительное число такое, что

$$1 + \epsilon_M \neq 1.$$

В машинной системе чисел с плавающей точкой $F = F(\beta, t, e_{min}, e_{max})$ любой вещественное число в диапазоне F может быть представлено с относительной ошибкой, не превосходящей ошибку округления u

$$u = \frac{1}{2}\beta^{-t+1}$$

(при использовании округления).

Стандарт IEEE 754 чисел с плавающей точкой

$$v = (-1)^s (1.m)_2 2^e \quad e_{min} \leq e \leq e_{max}.$$

Name	Common name	Base	Digits	E min	E max	Notes	Dec digits	Dec Emax
binary16	Half precision	2	10+1	−14	+15	storage	3.31	4.51
binary32	Single prec	2	23+1	−126	+127		7.22	38.23
binary64	Double prec	2	52+1	−1022	+1023		15.95	307.95
binary128	Quadruple pr	2	112+1	−16382	+16383		34.02	4931.77
decimal32		10	7	−95	+96	storage	7	96
decimal64		10	16	−383	+384		16	384
decimal128		10	34	−6143	+6144		34	6144

Математически эквивалентные формулы или алгоритмы не обязательно являются вычислительно эквивалентными.

Математическая эквивалентность – одинаковые результаты для одинаковых входных данных при использовании точной арифметики

Вычислительная эквивалентность – одинаковые (в пределах погрешности) результаты для одинаковых входных данных при использовании машинной арифметики

Пример: вычисление экспоненты от отрицательного аргумента.

$$\exp(-x) \approx \sum_{k=0}^{N} (-x)^k/k! \approx \left(\sum_{k=0}^{N} x^k/k! \right)^{-1}.$$

x=10, одинарная точность

$4.5399931\ 10^{-5}$ $-7.2657094\ 10^{-5}$ $4.5399924\ 10^{-5}$

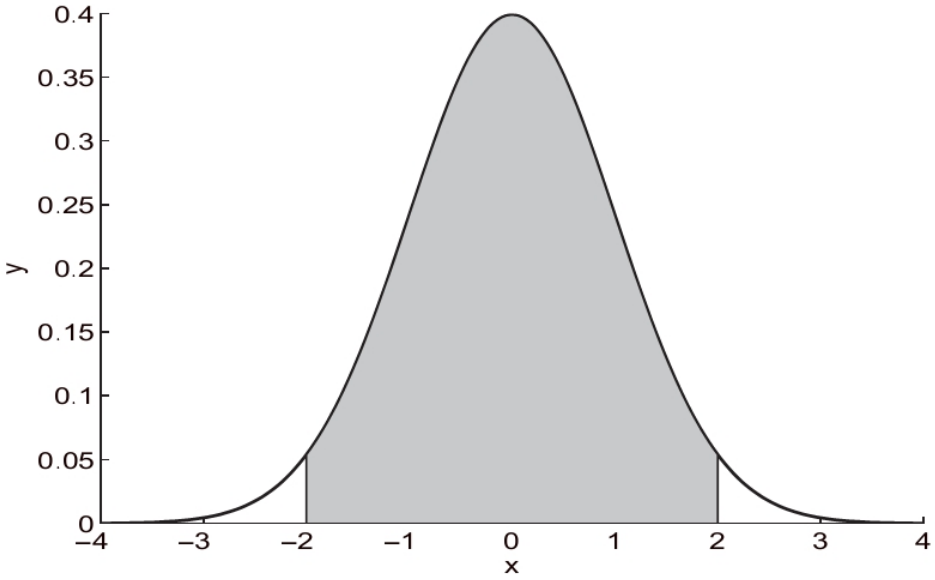
Статистическая модель ошибок округления

До сих пор обсуждались максимальные ошибки.
Они не учитывают знак ошибок (возможна компенсация!) и часто чересчур пессимистичны.
Альтернатива – статистический анализ в предположении, что ошибки независимы и случайны. (Хотя это выполняется не всегда.)

Если каждое значение x_i имеет погрешность $|\Delta_i| \leq \delta$, то максимальная оценка погрешности суммы $y = \sum_{i=1}^{n} x_i$
равна $n\delta$ – линейный рост.
Если числа округляются (а не усекаются!) и если предположить, что разные слагаемые статистически независимы с дисперсией ϵ, тогда стандартная ошибка y равна

$$(\epsilon^2 + \epsilon^2 + \ldots + \epsilon^2)^{1/2} = \epsilon\sqrt{n}$$

– пропорциональность корню из n.
Эмпирическое правило: если граница максимальной ошибки оценена как $f(n)u$, тогда можно ожидать фактическую ошибку $\sqrt{f(n)}u$.
(**Обязательна** случайность ошибок!).


0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
y
−4
−3
−2
−1
0
1
2
3
4
x

Переполнение и потеря точности

1) Переполнение в процессе вычислений – превышение максимальных машинных значений (Например, при вычислении длин векторов, модулей комплексных чисел и т.д.)

2) Потеря точности – существенное уменьшение числа значащих цифр в процессе вычислений (например, при вычитании двух близких чисел). Вычисление производных, приведение аргументов функций к удобным (малым) диапазонам.

Вычислительная задача: описание функциональной зависимости между входными данными (независимыми переменными) и выходными данными (желаемый результат).
Входные и выходные данные состоят из конечного числа вещественных (комплексных) величин.
Предполагается, что выходные данные однозначно определяются входными данными и непрерывно зависят от них.

Численный (вычислительный) метод: процедура, “аппроксимирующая” математическую задачу вычислительной задачей, или же решающая вычислительную задачу.

Распространение ошибок

В вычислительных задачах, как правило, входные данные неточны. В процессе вычислений эти погрешности эволюционируют и приводят к погрешностям результата.

Основные результаты о распространении ошибок

1) При сложении (вычитании):

$$y = \sum_{i=1}^{n} x_i, \qquad \Delta y \le \sum_{i=1}^{n} |\Delta x_i|$$

– граница абсолютной ошибки результата определяется суммой абсолютных ошибок данных.

2) При умножении (делении): то же верно для относительных ошибок:

$$y = \prod_{i=1}^{n} x_i^{m_i}, \qquad \left|\frac{\Delta y}{y}\right| \le \sum_{i=1}^{n} |m_i| \left|\frac{\Delta x_i}{x_i}\right|.$$

(Достаточно перейти к логарифмам).

Произвольная функция одного аргумента $y = f(x)$

Пусть $\Delta x = \tilde{x} - x$. Естественно аппроксимировать $\Delta y = \tilde{y} - y$ с помощью дифференциала y. По теореме о среднем,

$$\Delta y = f(x + \Delta x) - f(x) = f'(\xi)\Delta x, \quad \text{где} \quad \xi \in [x, x + \Delta x].$$

Пусть $|\Delta x| \leq \epsilon$. Тогда

$$|\Delta y| \leq \max_{\xi} |f'(\xi)|\epsilon, \quad \text{где} \quad \xi \in [x - \epsilon, x + \epsilon].$$

На практике ξ можно заменить на x, поскольку высокая относительная точность в оценке погрешности востребована редко.

Функция нескольких аргументов

Рассмотрим $f(x)$, $x = (x_1, x_2, \dots x_n)$, $\tilde{x} = x + \Delta x$. Тогда существует такое число θ, что

$$\Delta f = f(x + \Delta x) - f(x) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x + \theta \Delta x)\Delta x_i, \quad \text{где} \quad 0 \leq \theta \leq 1.$$

Доказательство следует из рассмотрения функции $F(t) = f(x + t\Delta x)$.

Функция нескольких аргументов

Таким образом, для дифференцируемой функции $f = f(x_1, x_2, \dots x_n)$ в окрестности $x = (x_1, x_2, \dots x_n)$ с погрешностями $\Delta x_1, \Delta x_2, \dots \Delta x_n$, погрешность

$$\Delta f \approx \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \Delta x_i.$$

Для максимальной погрешности верно

$$|\Delta f| \leq \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i} \right| |\Delta x_i|.$$

Строго говоря, в этой формуле должны браться максимумы частных производных в окрестности, однако часто берут значения в точке x.
(Это может быть неверно в экстремумах!).

Статистическая оценка

Полученная формула часто сильно переоценивает погрешность – помочь может статистическое рассмотрение.
Пусть погрешности $\Delta x_1, \Delta x_2, \dots \Delta x_n$ – независимые случайные переменные со средними нулевыми значениями и стандартными девиациями $\epsilon_1, \epsilon_2, \dots \epsilon_n$. Тогда стандартная погрешность ϵ для $f(x_1, x_2, \dots x_n)$ даётся формулой

$$\epsilon \approx \left(\sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i} \right)^2 \epsilon_i^2 \right)^{1/2}.$$

Плохо обусловленные задачи

Если “малые” изменения во входных данных приводят к “большим” изменениям в выходных данных, задача называет плохо обусловленной (иначе – хорошо обусловленной).

Рассмотрим вычислительную задачу $y = f(x) \in R^m$, $x \in R^n$. Зафиксируем $\hat{x}$ и предположим, что $\hat{x} \neq 0$ и $\hat{y} = f(\hat{x}) \neq 0$. Чувствительность y по отношению к малым изменениям в x может быть охарактеризована относительным числом обусловленности

$$\kappa(f; \hat{x}) = \lim_{\epsilon \to 0} \sup_{||h|| = \epsilon} \left\{ \frac{||f(x+h) - f(x)||}{||f(x)||} \Big/ \frac{||h||}{||x||} \right\}.$$

Для достаточно малых возмущений

$$||\tilde{y} - y|| \leq \kappa \epsilon ||y|| + O(\epsilon^2).$$

Системы линейных алгебраических уравнений (СЛАУ)

Многие задачи численного анализа сводятся к исследованию СЛАУ.
Как правило, система задаётся с помощью матрицы A размерности N.
Существуют различные задачи, два основных типа:

- Решение СЛАУ: матрица A и вектор b известны, требуется найти вектор x такой, что
$$Ax = b.$$
- Решение спектральной задачи (задачи на собственные значения): найти (все или некоторые) числа λ_i и вектора x_i такие, что
$$Ax_i = \lambda_i x_i.$$

Возможность и эффективность решения указанных задач определяются свойствами матрицы A.
В частности, с точки зрения вычислительной эффективности очень важно количество ненулевых элементов матрицы и их распределение в матрице.

- Плотно заполненные матрицы: количество ненулевых элементов $N_{nz} \sim N^2$
- Разреженные матрицы: количество ненулевых элементов $N_{nz} \ll N^2$
- Специальный тип разреженных матриц, ленточные матрицы: все ненулевые элементы расположены на нескольких ($\ll N$) субдиагоналях главной диагонали.

Элементы матрицы A могут как вычисляться заранее и храниться, так и вычисляться “на лету” при обращении к ним.

Решение СЛАУ: обусловленность

Задача $Ax = b$ должна быть корректной: решение должно существовать, быть единственным и непрерывно зависеть от входных данных.
Есть специфика вычислений: насколько большое изменение решения x вызывает изменение входных данных b?
Естественно оценивать относительную погрешность.
Пусть $A(x + \Delta x) = b + \Delta b$, $A\Delta x = \Delta b$. Оценим

$$\frac{||\Delta x||}{||x||} \Big/ \frac{||\Delta b||}{||b||} = \frac{||\Delta x||}{||\Delta b||} \frac{||b||}{||x||}.$$

Поскольку $\Delta x = A^{-1}\Delta b$, то $||\Delta x|| \le ||A^{-1}||\,||\Delta b||$. Далее,

$$||b|| \le ||A||\,||x||, \quad ||b||/||x|| \le ||A||.$$

Число обусловленности матрицы

Оценка чувствительности:

$$\frac{||\Delta x||}{||x||} \Big/ \frac{||\Delta b||}{||b||} \leq ||A||\ ||A^{-1}||.$$

Определение: Число обусловленности.

Пусть задана обратимая матрица A размерности N. Тогда

$$\kappa(A) = ||A||\ ||A^{-1}||,$$

где $||.||$ – некоторая матричная норма, называется числом обусловленности матрицы A по отношению к норме $||.||$.

Свойства $\kappa(A)$:

1) $\kappa(A) \geq 1$;

2) $\kappa(\alpha A) = \kappa(A)$.

Число обусловленности матрицы

Для вычисления числа обусловленности можно использовать, например, спектральную норму:

$$||A||_* = \max_{x \neq 0} \frac{||Ax||}{||x||},$$

где $||Ax||$ – Эвклидова норма вектора.

Спектральное число обусловленности,

$$\kappa^*(A) = ||A||_* \, ||A^{-1}||_* = \frac{\max_{\lambda \in \sigma(A)} |\lambda|}{\min_{\lambda \in \sigma(A)} |\lambda|},$$

обладает следующим минимальным свойством:

$$1 \leq \kappa^*(A) \leq \kappa(A),$$

где $\kappa(A)$ – число обусловленности для любой другой нормы.

Задача 1

Постройте график относительной погрешности при вычислении $\exp(-x)$ с помощью разложения в ряд Тейлора для подходящего диапазона x, $x \geq 0$. Сравните результаты для одинарной и двойной точности.

Литература

1. Dahlquist G., Bjoerck A., *Numerical Methods in Scientific Computing: Volume 1*, SIAM, 2008. Vol. 1. ISBN 0898716446. 793 p. Глава 2.