Численные методы

Василий Алфёров по лекциям Евгения Яревского

19 января 2019 г.

Содержание

1. Погрешности

Различают абсолютную и относительную погрешности.

Абсолютная погрешность – это модуль разности значения и экспериментальных данных. Обозначают её как ΔA или, иногда, ∇A.

Относительная погрешность – это $\frac{\Delta A}{A}$. Чем ближе к нулю, тем интереснее смотреть на относительную погрешность, а не на абсолютную.

Интересный пример большой относительной погрешности: после 1995 года масса нейтрино считалась как $-22 \pm 17_{stat} \pm 17_{syst}$ эВ^2. Сейчас уже измерили точнее (Нобелевская премия по физике 2015 года).

В английском языке погрешность обозначается словом “error”, имеющим нейтральную окраску. В русском языке слово “ошибка” имеет негативную окраску, поэтому чаще используют слово “погрешность”.

Источники погрешностей (ошибок):

A) Ошибки входных данных. Делятся на:

- Случайные. Подразумевают, что никакую величину в реальном мире нельзя измерить абсолютно точно.
- Систематические. Это либо приборные ошибки (подразумевают, что идеальных приборов не существует), либо погрешности в самом методе измерения.

B) Ошибки представления действительных чисел в компьютере и округления арифметических операций.

C) Ошибки из-за “обрезания” бесконечно малых и бесконечно больших величин.

D) Упрощения в математических моделях.

E) Человеческие и машинные ошибки. К машинным ошибкам можно отнести, например, историю с Pentium.

Ошибки типов A и D никак не контролируются, с ними приходиться смириться. Ошибки типа C, как правило, могут контролироваться. Ошибки типа B могут контролироваться частично.

Ошибки точно измерить нельзя, иначе бы они были не ошибками, а поправками. Поэтому обычно их оценивают сверху.

1.1. Представление действительных чисел в компьютере

Стандарт IEEE-754. Подразумевается, что мы всё это уже знаем, поэтому остановимся только на основных моментах.

Во-первых, мы можем сохранить в наших типах лишь конечное количество чисел. Из этого следует, например, что корректные с компьютерной точки зрения числа не образуют никакой алгебраической структуры. Если мы можем представить числа a и b, то мы не обязательно можем представить $a + b$. То же верно и для любой другой арифметической операции. Если $a + b = a$, то не обязательно $b = 0$. И у нас нет ни ассоциативности, ни дистрибутивности. Коммутативность, однако, есть.

Пример.
Иллюстрация отсутствия ассоциативности при одинарной точности (`float` в плюсах):

$$\sum_{n=1}^{10^9} \frac{1}{n} = 15.4036827087 \qquad \sum_{n=10^9}^{1} \frac{1}{n} = 18.8079185486$$

~~Не знаю, на каком пентиуме считались числа в презентации, у меня получилось вот так.~~ Настоящее значение равно при этом 21.3004815023. В реальной жизни используются, в основном, числа двойной точности (`double` в плюсах и джаве, `float` в питоне). Видимо, по поводу того, что `long double` из коробки есть только в плюсах, в остальных языках за ним надо лезть в неочевидные библиотеки.

***Определение* 1.1.**
Математически эквивалентные алгоритмы – алгоритмы, эквивалентные в предположении, что используется точная арифметика.

***Определение* 1.2.**
Вычислительно эквивалентные алгоритмы – алгоритмы, эквивалентные при использовании машинной арифметики с небольшой погрешностью.

Это не одно и то же.

Пример.
Вычислим e^{-10}, используя одинарную точность, двумя способами:

$$e^{-10} = \sum_{k=0}^{N} \frac{(-10)^k}{k!} = -6.25618267804 \cdot 10^{-5} \qquad e^{-10} = \left(\sum_{k=0}^{N} \frac{10^k}{k!} \right)^{-1} = 4.5399923671 \cdot 10^{-5}$$

Как видно, вычислительно способы не эквивалентны, хотя математически оба ряда сходятся к e^{-10}. Настоящее значение равно $4.53999297624849 \cdot 10^{-5}$. Очень большая погрешность в первом способе объясняется тем, что в начале мы попеременно складываем положительные и отрицательные слагаемые, далёкие по модулю от нуля.

Иногда, чтобы представлять порядок ошибки, используют интервальную арифметику.

1.2. Статистические оценки погрешности

Максимальные оценки погрешности зачастую пессимистичны, ведь они не учитывают знак. Обычно всё же ошибки друг друга компенсируют. Альтернативой является статистический анализ.

В рамках статистического анализа обычно считается, что ошибки независимы и случайны, хотя это выполняется и не всегда.

Пример.
Пусть каждое значение x_i имеет погрешность $|\Delta x_i| \leqslant \delta$. Тогда максимальная погрешность их суммы $y = \sum x_i$ оценивается как

$$|\Delta y| \leqslant \sum_{i=1}^{n} |\Delta x_i| \leqslant n\delta$$

Пусть теперь числа при операциях округляются (не усекаются, то есть нету перекоса от округления вниз и матожидание ошибок равно нулю). Пусть также мы считаем, что дисперсия

каждой из ошибок x_i ограничена сверху числом ε. Тогда дисперсия ошибки их суммы будет оценена как

$$D[\Delta y] \leqslant \sqrt{\sum_{i=1}^{n} \varepsilon^2} = \varepsilon\sqrt{n}$$

Эмпирически хорошо работает правило: если максимальная ошибка оценивается как $uf(n)$, то дисперсия будет оцениваться как $u\sqrt{f(n)}$. Для того, чтобы это работало, обязательно требуется, чтобы матожидание ошибки было нулевым.

1.3. Переполнение и потеря точности

Переполнение – превышение максимальных допустимых значений (на минуточку, у `double` это порядка $1.8 \cdot 10^{308}$). Возникает, например, при попытке вычислить модуль вектора или комплексного числа~~, когда оба компонента комплексного числа или вектора имеют порядок 10^{154}, видимо~~. Предлагаемый способ борьбы: заранее вынести большую константу из-под корня.

Потеря точности – существенное уменьшение числа значащих цифр в процессе вычислений. Например, возникает при вычитании близких больших чисел. Способы борьбы: считать производные или приводить аргументы функций к малым диапазонам.

1.4. Распространение ошибок

Входные данные, как мы уже выяснили, в вычислительных задачах, как правило, неточные. В ходе вычислений их погрешности эволюционируют и приводят к погрешности результата. В этом разделе мы обсудим конкретные оценки ошибок.

Теорема 1.1 (Сложение и вычитание)**.**
Пусть у величин x_1, ..., x_n известны максимальные погрешности $|\Delta x_1|$, ..., $|\Delta x_n|$.

Тогда у величины $y = \sum\limits_{i=1}^{n} x_i$ максимальная погрешность оценивается как $|\Delta y| \leqslant \sum\limits_{i=1}^{n} |\Delta x_i|$.

Доказательство.
Побуду занудой и скажу, что это неравенство треугольника для модуля. □

Теорема 1.2 (Произвольная функция одного аргумента)**.**
Пусть у величины x известная максимальная погрешность $|\Delta x|$.

Пусть также $f \in C^1[x, x + \Delta x]$.

Обозначим величину $y = f(x)$.

Тогда существует такое $\xi \in [x, x + \Delta x]$, что $|\Delta y| \leqslant |f'(\xi)\Delta x|$.

Доказательство.
По теореме Лагранжа о среднем значении, существует ξ такое, что $f'(\xi)\Delta x = f(x+\Delta x)-f(x)$. Остаётся лишь заметить, что $\Delta y = |f(x + \Delta x) - f(x)|$. □

На практике часто считают Δx достаточно маленьким и берут $\xi = x$. Однако это не работает в случае, если у f в точке x экстремум – тогда нужно писать более строгие оценки.

Теорема 1.3 (Умножение и деление)**.**
Пусть у величин x_1, ..., x_n известны максимальные относительные погрешности $|\frac{\Delta x_1}{x_1}|$, ..., $|\frac{\Delta x_n}{x_n}|$.

Тогда у величины $y = \prod\limits_{i=1}^{n} x_i^{m_i}$ максимальная относительная погрешность оценивается как $|\frac{\Delta y}{y}| \leqslant \sum\limits_{i=1}^{n} |m_i||\frac{\Delta x_i}{x_i}|$.

Доказательство.

$$\left|\frac{\Delta y}{y}\right| = |\ln' y \cdot \Delta y| \leqslant |\Delta \ln y| = \left|\Delta\left(\sum_{i=1}^{n} m_i \ln x_i\right)\right| \leqslant \sum_{i=1}^{n} m_i |\Delta \ln x_i| \leqslant \sum_{i=1}^{n} m_i \left|\frac{\Delta x_i}{x_i}\right|$$

□

Теорема 1.4 (Функция нескольких переменных)**.**
Пусть у величин x_1, ..., x_n известны максимальные погрешности Δx_1, ..., Δx_n.

Пусть также $f : \mathbb{R}^n \to \mathbb{R}^n$ – функция, непрерывно дифференцируемая на отрезке $[x; x + \Delta x]$.

Обозначим величину $y = f(x_1, \ldots, x_n)$.

Тогда существует такое $\theta \in [0, 1]$, что $|\Delta y| \leqslant \sum\limits_{i=1}^{n} \left|\frac{\partial f}{\partial x_i}(x + \theta\Delta x)\right| |\Delta x_i|$.

Доказательство.
Применим теорему 1.2 к функции $F(t) := f(x + t\Delta x)$. □

Опять же, нередко берут значения частных производных в точке x, что точно так же может оказаться неверным в экстремумах.

Статистическая погрешность в последнем случае оценивается как

$$\varepsilon \approx \sqrt{\sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i}\right)^2 \varepsilon_i^2}$$

1.5. Плохо обусловленные задачи

Если маленькие изменения во входных данных вызывают большие изменения в выходных данных, то задачу называют плохо обусловленной, иначе – хорошо обусловленной. Чем хуже обусловлена задача, тем большие требования по погрешности предъявляются к её решениям.

***Определение* 1.3.** Рассмотрим вычислительную задачу $f(x) : \mathbb{R}^n \to \mathbb{R}^m$.

Зафиксируем входные данные $\hat{x} \neq 0$ и решение $\hat{y} = f(\hat{x}) \neq 0$.

Относительным числом обусловленности мы в таком случае назовём число

$$\kappa(f, \hat{x}) = \lim_{\varepsilon \to 0} \sup_{\|h\| = \varepsilon} \left\{ \frac{\|f(x + h) - f(x)\|}{\|f(x)\|} \cdot \frac{\|x\|}{\|h\|} \right\}$$

Формально, из этого определения получится, что для достаточно малых возмущений (норма которых ограничена ε) будет выполняться

$$\|\hat{y} - y\| \leqslant \kappa\varepsilon\|y\| + (\varepsilon^2)$$

2. Ряды. Суммирование и ускорение сходимости

TODO: