

1 Основные понятия вычислительной математики

1.1 Предмет и особенности вычислительной математики

Вычислительная математика — раздел математики, включающий круг вопросов, связанных с производством вычислений и использованием компьютеров. В более узком понимании вычислительная математика — теория численных методов решения типовых математических задач.

К задачам вычислительной математики относят:

- решение систем линейных уравнений;
- решение систем нелинейных алгебраических уравнений;
- нахождение собственных значений и векторов матрицы;
- решение систем дифференциальных уравнений;
- решение интегральных уравнений;
- задачи аппроксимации;
- задачи интерполяции;
- задачи экстраполяции;
- задачи численной оптимизации.

Вычислительная математика (численный анализ) имеет отличия от других дисциплин и обладает определенными особенностями:

- Она имеет дело как с дискретными, так и непрерывными объектами. Например, вместо отрезка прямой рассматривается набор точек, вместо непрерывной функции — табличная функция, вместо производных — их разностные аппроксимации. Дискретизация является необходимой для переноса математической модели в вычислительную среду. Таким образом, уже на этапе постановки задачи возникает некоторая погрешность.
- Численный анализ сталкивается с двумя типами ошибок: **устраняемыми** и **неустраняемыми**. При этом и устранимость ошибки лишь потенциальная (например, устраняемая ошибка округления исчезает, если вычислительная машина является абсолютно точной). Неустраняемая ошибка, как правило, заложена в самом вычислительном методе, математической модели реального объекта и входных данных.
- Задачи вычислительной математики обладают особым качеством — **обусловленностью**. Как правило, под обусловленность задачи понимается чувствительность ее решения к малым возмущениям входных данных.
- При решении вычислительных задач выбор метода является часто главным фактором, влияющим на качество решения.
- Для вычислительной математики важным фактором, влияющим на выбор способа решения задачи, является **экономичность**, т.е. стремление минимизировать вычислительные затраты на решение.

Рассмотрим некоторые примеры, иллюстрирующие эти особенности.

1.2 Примеры

1.2.1 Обусловленность задачи

Пример. Вычислить все корни уравнения

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

Точное решение легко можно найти: $x_1 = 2.01, x_2 = 1.99, x_{3,4} = 2 \pm 0.01i$. Однако, если точность вычислений больше, чем 10^{-8} , то с численной точки зрения из-за округления свободного члена решается уравнение $(x - 2)^4 = 0$, корни которого $x_{2,3,4} = 2$.

Пример. Пусть решается задача Коши для ОДУ 2-го порядка:

$$y''(x) = y(x), \quad y(0) = 1, \quad y'(0) = -1.$$

Общее решение имеет вид

$$y(x) = \frac{(y(0) + y'(0))e^x + (y(0) - y'(0))e^{-x}}{2}.$$

При заданных выше начальных условиях точное решение задачи $y(x) = e^{-x}$. Однако, если начальные условия заданы с малой ошибкой δ , в решении появляется быстрорастущее слагаемое δe^x , которое искажает результат.

Пример. Пусть решается задача Коши:

$$y' = 10y, \quad y(x_0) = y_0, \quad x \in [0; 1].$$

Общее решение имеет вид $y(x) = y_0 e^{10(x-x_0)}$. Пусть начальное условие известно лишь приближенно, т.е. $y(x_0) \approx y_0^*$. Решая эту приближенную задачу, мы найдем лишь $y^*(x) = y_0^* e^{10(x-x_0)}$, которое отличается от истинного решения на величину

$$y^* - y = (y_0^* - y_0) e^{10(x-x_0)}.$$

Пусть требуется обеспечить точность решения на всем промежутке не меньше заданного положительного ε . Нетрудно оценить

$$\max_{x \in [0; 1]} |y^*(x) - y(x)| = |y^*(1) - y(1)| = |y_0^* - y_0| e^{10(1-x_0)}.$$

Очевидно, требуемую точность решения можно получить, если точность задания начального условия $\delta \leq \varepsilon e^{-10}$.

Таким образом, ограничения на точность начальных данных в e^{10} раз выше требуемой точности решения и поэтому абсурдны.

Пример. Пусть решается система линейных алгебраических уравнений (СЛАУ) вида

$$\begin{cases} x_1 + 10x_2 = 11 \\ 100x_1 + 1001x_2 = 1101 \end{cases}.$$

Легко проверить, что решение системы является пара $(1, 1)$. В то же время возмущенная система

$$\begin{cases} x_1 + 10x_2 = 11.01 \\ 100x_1 + 1001x_2 = 1101 \end{cases}.$$

имеет решение $(11.01, 0)$, сильно отличающееся от решения невозмущенной.

Пример. Рассмотрим классический пример с многочленом

$$P(x) = (x - 1)(x - 2) \dots (x - 20).$$

Если коэффициент при x^{19} в канонической записи $P(x)$ изменить на величину порядка 10^{-7} , корни многочлена очень сильно изменяться (в частности, появятся комплексные).

1.2.2 Влияние выбора вычислительного метода

Пример. Из анализа известно, что функция синус разлагается в ряд Тейлора, который сходится при любых значениях аргумента:

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots$$

Вычислим по этой формуле синус 30° , т.е. с точностью до четвертого знака, $x = 0.5236$. Суммируя члены ряда, большие 10^{-4} , получим $\sin(0.5236) = 0.5000$.

Пусть теперь $x = 25.66$ (1470°). Если вычисления производить с восемью значащими цифрами, получим $\sin(25.25) \approx 24$, что невозможно по определению функции синус.

Как можно исправить ситуацию?

Пример. Рассмотрим достаточно известный интеграл:

$$I_n = \int_0^1 x^n e^{1-x} dx, \quad n = 1, 2, \dots$$

Интегрируя по частям, легко можно получить следующую рекуррентную формулу:

$$I_0 = e - 1, \quad I_n = nI_{n-1} - 1, \quad n \geq 1.$$

Пусть в вычислениях мы значение I_0 округлили до шести значащих цифр. Легко проверить, что уже I_9 получится при этом отрицательным, что невозможно.

Где накапливается погрешность?

Пример. Рассмотрим рекуррентное соотношение:

$$x_i = qx_{i-1}, \quad i \geq 0, \quad x_0 = a > 0, \quad q > 0.$$

Пусть при вычислениях из-за конечности разрядной сетки на i -ом шаге возникла ошибка округления и «машинное» решение исказилось: $x_i^M = x_i + \delta$. Тогда на следующем шаге

$$x_{i+1}^M = q(x_i + \delta) = qx_i + q\delta = x_{i+1} + q\delta.$$

Очевидно, если $|q| > 1$, то погрешность, возникающая из-за **одной** ошибки округления, будет расти. Заметим, что в этом случае говорят, что алгоритм **неустойчив**.

1.2.3 Экономичность метода

Пример. Классический пример — вычисление значений многочленов. Если вычислять значение

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0$$

в некоторой точке, подчиняясь стандартной форме записи многочлена, то нужно будет выполнить $n^2 + \left\lceil \frac{n}{2} \right\rceil$ умножений и n сложений.

Однако еще в средние века китайским математикам был известен метод, который сейчас называется обычно схемой Горнера:

$$P(x) = ((\dots (a_n x + a_{n-1}) x + a_{n-2}) x + \dots + a_0).$$

Этот метод требует всего n умножений и столько же сложений.

Пример. Решение СЛАУ с трехдиагональной матрицей каким-нибудь общим методом (методом Гаусса, например), требует порядка n^3 арифметических действий. Учет структуры матрицы позволяет число операций уменьшить до n (метод прогонки).

1.2.4 Погрешность метода

Воспользовавшись определением производной, запишем формулу приближенного вычисления $f'(x)$:

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

Чтобы оценить погрешность при вычислении по такой формуле, разложим функцию в ряд Тейлора и получим

$$\frac{f(x+h) - f(x)}{h} = \frac{[f(x) + hf'(x) + O(h^2)] - f(x)}{h} = f'(x) + O(h).$$

Таким образом, погрешность метода есть $O(h)$.

Если же повторить процедуру для следующей формулы:

$$f'(x) \approx \frac{f(x+h) - f(x-h)}{2h},$$

то окажется, что погрешность метода $O(h^2)$. Заметим, что стремится сделать шаг h как можно меньше бессмысленно. Легко можно показать, что существует некоторое оптимальное значение h^* такое, что при дальнейшем уменьшении шага погрешность вычислений начнет расти за счет ошибок округления, хотя погрешность метода будет все меньше.

Докажите это и найдите h^* !

1.3 Элементы теории погрешностей

Определение. Пусть u — точное значение некоторой величины, а u^* — приближенное. Абсолютной погрешностью приближения u^* называется величина $\Delta(u^*)$, удовлетворяющая неравенству $|u - u^*| \leq \Delta(u^*)$.

Определение. Относительной погрешностью приближения u^* называется величина $\delta(u^*)$, удовлетворяющая неравенству $\left| \frac{u - u^*}{u^*} \right| \leq \delta(u^*)$.

Определение. Пусть $u = u(t_1, t_2, \dots, t_n)$ определена на множестве параметров Ω . Предельной абсолютной погрешностью называется величина

$$D(u^*) = \sup_{(t_1, \dots, t_n) \in \Omega} |u(t_1, \dots, t_n) - u^*|.$$

Определение. Пусть $u = u(t_1, t_2, \dots, t_n)$ определена на множестве параметров Ω . Предельной относительной погрешностью называется величина $\frac{D(u^*)}{|u^*|}$.

Чтобы сформулировать следующее определение, необходимо напомнить один полезный факт из анализа.

Утверждение 1.1 Если функция $f(x) = f(x_1, \dots, x_n)$ определена и дифференцируема в каждой точке области Ω , то для каждой пары точек $x = (x_1, \dots, x_n) \in \Omega$ и $x + \Delta x = (x_1 + \Delta x_1, \dots, x_n + \Delta x_n) \in \Omega$ найдется такая точка $\xi = (\xi_1, \dots, \xi_n)$, лежащая на отрезке с концами в x и $x + \Delta x$, что

$$f(x + \Delta x) - f(x) = \sum_{i=1}^n \frac{\partial f(\xi)}{\partial x_i} \Delta x_i.$$

Это выражение называется формулой конечных приращений (Лагранжа).

Пусть теперь $u^* = u(t_1^*, \dots, t_n^*)$, причем $|t_i - t_i^*| \leq \Delta(t_i^*)$, $i = 1, \dots, n$. Предполагая, что функция u непрерывно дифференцируема, по формуле Лагранжа получаем

$$u(t_1, \dots, t_n) - u^* = \sum_{i=1}^n u'_{t_i}(t_1^* + \alpha(t_1 - t_1^*), \dots, t_n^* + \alpha(t_n - t_n^*))(t_i - t_i^*),$$

где $0 \leq \alpha \leq 1$. Если обозначить $\beta_i = \sup_{\Omega} |u'_{t_i}(t_1, \dots, t_n)|$, то получаем оценку

$$|u(t_1, \dots, t_n) - u^*| \leq D_1(u^*) = \sum_{i=1}^n \beta_i \Delta(t_i^*).$$

На практике используют и более грубую, линейную оценку погрешности:

$$|u(t_1, \dots, t_n) - u^*| \leq D_2(u^*) = \sum_{i=1}^n |u'_{t_i}(t_1^*, \dots, t_n^*)| \Delta(t_i^*).$$

Действительно, в силу непрерывности частных производных,

$$u'_{t_i}(t_1^* + \alpha(t_1 - t_1^*), \dots, t_n^* + \alpha(t_n - t_n^*)) = u'_{t_i}(t_1^*, \dots, t_n^*) + o(1).$$

Поэтому $\beta_i = |u'_{t_i}(t_1^*, \dots, t_n^*)| + o(1)$, откуда и следует требуемая оценка (вообще говоря, неверная).

Можно показать, что

1. Предельная абсолютная погрешность суммы или разности равна сумме предельных погрешностей.
2. Предельная относительная погрешность произведения или частного приблизительно равна сумме предельных относительных погрешностей.

Рассмотрим несколько примеров.

Пример. Пусть $u(t) = t^{10}$, $t^* = 1$, $\Delta(t^*) = 0.001$. Найдем $D(t^*)$, $D_1(t^*)$, $D_2(t^*)$.

$$\begin{aligned} u^* &= (t^*)^{10} = 1, & u'_t &= 10t^9, \\ u'_t(t^* + \alpha(t - t^*)) &= 10(1 + \alpha(t - 1)), \\ \beta &= \sup_{|t-1| \leq 0.001} |10t^9| = 10.09 \dots \end{aligned}$$

Теперь легко вычислить

$$\begin{aligned} D(u^*) &= \sup_{|t-1| \leq 0.001} |t^{10} - 1| = 0.010045 \dots, \\ D_1(u^*) &= \beta \Delta(t^*) = 0.01009 \dots, \\ D_2(u^*) &= |u'_t(t^*)| \Delta(t^*) = 0.01 \dots \end{aligned}$$

Как видим, в этом случае точная и линейная оценки погрешности не сильно отличаются.

Пример. Пусть теперь $\Delta(t^*) = 0.1$. В этом случае $\beta = \sup_{|t-1| \leq 0.1} |10t^9| = 23.5 \dots$ и получаем

$$\begin{aligned} D(u^*) &= \sup_{|t-1| \leq 0.1} |t^{10} - 1| = 1.59 \dots, \\ D_1(u^*) &= \beta \Delta(t^*) = 2.35 \dots, \\ D_2(u^*) &= |u'_t(t^*)| \Delta(t^*) = 1 \dots \end{aligned}$$

Здесь разница между оценками более заметна.

Пример. Пусть $F(x, p, q) = x^2 + px + q$. Оценим погрешность корней уравнения $F(x, p, q) = 0$ при заданных приближенных значениях коэффициентов p^*, q^* и их погрешностях $\Delta(p^*), \Delta(q^*)$.

Пусть x^* — корень уравнения $F(x, p^*, q^*) = 0$. Чтобы найти предельные погрешности, необходимо знать x'_p и x'_q . Дифференцируя уравнение относительно коэффициентов и разрешая получившуюся систему, получаем, что

$$\begin{aligned}\frac{\partial x}{\partial p} &= -\frac{\partial F}{\partial p} \left(\frac{\partial F}{\partial x} \right)^{-1}, \\ \frac{\partial x}{\partial q} &= -\frac{\partial F}{\partial q} \left(\frac{\partial F}{\partial x} \right)^{-1}.\end{aligned}$$

Итак, $x'_p = -\frac{x}{2x+p}$ и $x'_q = -\frac{1}{2x+p}$. Тогда линейная оценка погрешности имеет вид

$$D_2(x^*) = \frac{|x^*| \Delta(p^*)}{|2x^* + p^*|} + \frac{\Delta(q^*)}{|2x^* + p^*|}.$$

Очевидно, что $D_2(x^*) \rightarrow \infty$ при $2x^* + p^* \rightarrow 0$, т.е. в некоторых случаях оценка может быть очень завышена.

Задача 1.3.1. С каким числом знаков надо взять $\lg 2$, чтобы вычислить корни уравнения $x^2 - 2x + \lg 2 = 0$ с четырьмя верными знаками?

Задача 1.3.2. Найти относительную погрешность $\delta(u)$ вычисления функции $u = xy^2z^3$ в точке $x^* = 37.1, y^* = 9.87, z^* = 6.052$, если известны погрешности координат $\Delta(x^*) = 0.3, \Delta(y^*) = 0.11, \Delta(z^*) = 0.016$.

Задача 1.3.3. Вычислить относительную погрешность в определении значения функции $u = \frac{x^2y^2}{z^4}$, если заданы $x^* = 37.1, y^* = 9.87, z^* = 6.052$ с известными погрешностями $\Delta(x^*) = 0.1, \Delta(y^*) = 0.05, \Delta(z^*) = 0.02$.

Задача 1.3.4. Найти абсолютную предельную погрешность $D(u^*)$, погрешность по производной $D_1(u^*)$ и линейную оценку погрешности $D_2(u^*)$ для функций $u = \sin t, u = \frac{1}{t^2 - 5t + 6}$. Заданы точка приближения $t = t^*$ и погрешность $\Delta(t)$.

Задача 1.3.5. Пусть для вычисления функции $u(t)$ используется частичная сумма ряда Маклорена

$$u(t) \approx u(0) + \frac{u'(0)}{1!}t + \frac{u''(0)}{2!}t^2 + \dots + \frac{u^{(n)}(0)}{n!}t^n,$$

причем аргумент задан с погрешностью $\Delta t = 10^{-3}$.

Найти n такое, чтобы погрешность определения $u(t)$ не превышала Δt . Рассмотреть отрезки $t \in [0; 1]$ и $t \in [10; 11]$. Оценить число слагаемых для функций $u = \cos t$ и $u = e^t$ на втором из отрезков.

2 Близость и нормы

Рассмотрим вопрос численного нахождения интеграла $I = \int_a^b x(t) dt$. Можно заменить подынтегральную функцию на более простую, интеграл от которой легко вычисляется. Главное, чтобы она была достаточно «близка» к исходной. А можно заменить интеграл суммой $\sum_i x(t_i) \Delta t_i$, которая тоже просто вычисляется.

В первом случае приближенный метод вычисления основан на замене исходных данных близкими к ним. Во втором случае заменяется на более простой сам оператор интегрирования. Очевидно, под близостью в этих двух случаях понимаются разные вещи. Формализация этого интуитивного понятия составляет важную часть функционального анализа. Напомним некоторые понятия, которые будут в дальнейшем часто использоваться.

Определение. Метрическим пространством называется множество элементов x произвольной природы, на котором введено расстояние (метрика) $\rho(x_1, x_2)$ между любыми двумя элементами множества, удовлетворяющее следующим условиям:

1. $\rho(x_1, x_2) \geq 0$.
2. $\rho(x_1, x_2) = 0 \Leftrightarrow x_1 = x_2$.
3. $\rho(x_1, x_2) = \rho(x_2, x_1)$.
4. $\rho(x_1, x_3) \leq \rho(x_1, x_2) + \rho(x_2, x_3)$.

Определение. Последовательность элементов x_n метрического пространства называется сходящейся (по метрике) к элементу x^* , если $\lim_{n \rightarrow \infty} \rho(x_n, x^*) = 0$.

Определение. Последовательность элементов x_n метрического пространства называется фундаментальной, если для любого $\varepsilon > 0$ найдется такой номер $k = k(\varepsilon)$, что $\rho(x_n, x_m) < \varepsilon$ при всех n и m больших k .

Определение. Метрическое пространство называется полным, если любая фундаментальная последовательность его элементов сходится к элементу того же пространства.

Последнее качество очень важно для обоснования вычислительных методов. Действительно, если некоторый итерационный метод имеет дело с элементами неполного пространства (например, множеством рациональных чисел), то невозможно обосновать его сходимости: даже если последовательность приближений является фундаментальной, то она не обязательно сходится к элементу этого же пространства, т.е. приближенное решение будет недопустимым с точки зрения постановки задачи.

Определение. Линейное или векторное пространство L над полем P — это множество, на котором введены операции сложения и умножения на скаляр (элемент поля P), удовлетворяющие следующим условиям:

1. $(L, +)$ — абелева группа.
2. Операция умножения на скаляр ассоциативна.
3. Умножение на единицу поля P сохраняет элемент пространства неизменным.

4. Операция умножения скаляра на элемент L дистрибутивна относительно сложения скаляров.
5. Операция умножения элемента L на скаляр дистрибутивна относительно сложения элементов пространства L .

Определение. Скалярным произведением в векторном пространстве L над полем \mathbb{C} называется функция (x, y) для элементов (векторов) $x, y \in L$, принимающая значения в \mathbb{C} и удовлетворяющая следующим условиям:

1. $(\alpha x_1 + \beta x_2, y) = \alpha(x_1, y) + \beta(x_2, y)$.
2. $(x, y) = \overline{(y, x)}$.
3. $(x, x) \geq 0$, причем $(x, x) = 0 \Leftrightarrow x = 0$.

Определение. Нормой в векторном пространстве L над полем вещественных или комплексных чисел называется отображение $\|\cdot\| : L \rightarrow \mathbb{R}^+$, обладающее следующими свойствами:

1. $\forall x \in L, \|x\| \geq 0$.
2. $\|x\| = 0 \Leftrightarrow x = 0$.
3. $\forall x, y \in L, \|x + y\| \leq \|x\| + \|y\|$ (неравенство треугольника).
4. $\forall \alpha \in \mathbb{R}, \forall x \in L, \|\alpha x\| = |\alpha| \|x\|$.

Векторное пространство с введенной на нем нормой называется нормированным пространством.

В любом нормированном пространстве можно ввести метрику, определив ее как $\rho(x, y) = \|y - x\|$ (легко проверить, что при этом выполняются все аксиомы метрики). В этом случае говорят, что метрика порождена нормой.

Определение. Банахово пространство — нормированное векторное пространство, полное по метрике, порожденной нормой пространства.

Вычислительная математика чаще всего имеет дело со следующими банаховыми пространствами:

1. Множество вещественных чисел \mathbb{R} с нормой $\|x\| = |x|$.
2. Множество $C[a, b]$ функций, непрерывных на отрезке $[a, b]$ с чебышевской нормой $\|x\|_c = \max_{t \in [a, b]} |x(t)|$. Заметим, что сходимость по метрике, порожденной этой нормой, называется равномерной.
3. Множество $L_p[a, b]$ функций $x(t)$, определенных на отрезке $[a, b]$ и интегрируемых по модулю с p -ой степенью. Норма задается формулой

$$\|x\|_{L_p} = \left(\int_a^b |x(t)|^p dt \right)^{\frac{1}{p}}.$$

Сходимость в этой норме называется сходимостью в среднем. В отличие от равномерной близости, функции, близкие в среднем, могут мало отличаться на большей

части области определения и сильно — на небольших участках. Выбор пространства и нормы (а значит, и метрики) обусловлен «физическими» соображениями, т.е. диктуется ограничениями прикладной области задачи.

Можно показать, что между различными нормами существует определенная связь:

$$\|x\|_{L_1} \leq \|x\|_{L_2} \leq \dots \leq \|x\|_c.$$

Таким образом, чебышевская норма оказывается самой сильной: из сходимости в этой норме следует сходимость и в любой норме $\|\cdot\|_p$.

4. Конечномерное пространство \mathbb{R}^n векторов $x = (x_1, x_2, \dots, x_n)$. Норму вектора можно ввести разными способами. Довольно часто выбирают одну из гёльдеровых норм: $\|x\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}}$. Частными ее случаями являются

$$\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2} = \sqrt{(x, x)} \quad \text{— евклидова норма,}$$

$$\|x\|_\infty = \max_i |x_i| \quad (p \rightarrow \infty).$$

5. Пространство квадратных матриц порядка n . Нормирование этого пространства обычно привязывают к нормированию пространства векторов.

Норма матрицы называется согласованной с нормой вектора, если для любой матрицы A и любого вектора x справедливо $\|Ax\| \leq \|A\| \cdot \|x\|$. Говорят, что норма матрицы подчинена норме вектора, если

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

В соответствии с этим, рассматривают нормы матриц

$$\|A\|_\infty = \max_i \sum_j |a_{ij}|,$$

$$\|A\|_1 = \max_j \sum_i |a_{ij}|,$$

$$\|A\|_2 = \sqrt{\max_i \lambda_{A^*A}^i}.$$

Здесь λ_D^i — собственные числа матрицы D . Последняя норма также называется спектральной.

Задача 2.0.6. Построить примеры последовательностей функций, которые сходятся в смысле чебышевской нормы, нормы пространства L_p .

Задача 2.0.7. Пусть функция $f(x)$ задана таблично. Написать на языке \mathcal{N} функцию, которая бы приближенно вычисляла нормы $\|\cdot\|_c$ и $\|\cdot\|_{L_p}$. Найти готовые реализации (библиотеки, пакеты и проч.) для языка \mathcal{N} .

Задача 2.0.8. Доказать подчиненность перечисленных выше норм матриц соответствующим векторным нормам.

Задача 2.0.9. Написать на языке \mathcal{N} функцию, которая бы вычисляла нормы $\|\cdot\|_\infty$ и $\|\cdot\|_p$ для векторов и матриц. Найти готовые реализации (библиотеки, пакеты и проч.) для языка \mathcal{N} .