Численное дифференцирование. Погрешность метода. Вычислительная погрешность.

До сих пор рассматривались вопросы, относящиеся к обусловленности задачи и экономичности численного метода. Не смотря на важность этих вопрос, ключевое место в вычислительной математике занимает изучение погрешности численного метода и вычислительной погрешности. Каждый раздел математики имеет свою специфику в разрешении этого вопроса. Первое знакомство с погрешностью метода и вычислительной погрешностью проведем на примере численного дифференцирования.

Напомним, что по определению производная

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}. \tag{2.1}$$

Численное вычисление производной подразумевает переход от непрерывных функций к дискретным или сеточным (табличным) функциям. Для этого вводят сетку как множество дискретизации значений. Рассмотрим равномерную сетку с шагом h. Тогда на отрезке $[a,b]$ сетка задается следующим образом

$$x_i = a + ih,\ i = 0,...,N, \tag{2.2}$$

где $N = \frac{b-a}{h}$. Пусть $f(x_i) = f_i$ — функция, определенная в узлах сетки.

Предложим два способа вычисления производной. Первый способ

$$f'(x_i) \approx \frac{f(x_i+h) - f(x_i)}{h}, \tag{2.3}$$

второй способ (схема с центральной точкой)

$$f'(x_i) \approx \frac{f(x_i+h) - f(x_i-h)}{2h}. \tag{2.4}$$

Разумеется это приближенное вычисление производной. Оценим погрешность этих методов. Для этого введем численную производную следующим образом

$$f_i' = \frac{f_{i+1} - f_i}{h} \text{ и } f_i' = \frac{f_{i+1} - f_{i-1}}{2h}$$

соответственно.

Основным механизмом оценки погрешности метода является разложение в ряд Тейлора вокруг требуемой точки. Поскольку в обоих случаях требуемой точкой является x_i, то именно вокруг нее и будем осуществлять разложение. Для $f(x_i+h)$ имеем

$$f(x_i+h) = f(x_i) + f'(x_i)h + \frac{f''(x_i)}{2!}h^2 + O(h^3). \tag{2.5}$$

Следует обратить внимание на то, что в (2.5) под $f(x_i)$ подразумевается некоторая аналитическая функция, которая имеет как минимум 3 непрерывных производных. Однако (2.5) можно спроецировать на сетку (2.2) и получить следующее соотношение

$$[f]_{i+1} = [f]_i + [f']_i h + \frac{[f'']_i}{2!}h^2 + O(h^3), \tag{2.6}$$

где квадратные скобки “[]” информируют о проекции функции на сетку. Более подробно об особенностях разложения в ряд будет рассмотрено в параграфе, посвященному численному решению систем обыкновенных дифференциальных уравнений. Однако уже на текущем этапе следует понимать, что разложение сеточной функции в ряд как токовой невозможно. На практике для экономии времени и места квадратные скобки опускаются. Но для четкого понимания происходящего далее квадратные скобки будем оставлять.

Подставляем (2.6) в выражение для погрешности

$$f_i' = \frac{[f]_{i+1} - [f]_i}{h} = \frac{[f]_i + [f']_i h + \frac{[f'']_i}{2!} h^2 + 0(h^3) - [f]_i}{h} = [f']_i + \frac{[f'']_i}{2} h + O(h^2).$$

Погрешность метода для (2.3)

$$f_i' - f'(x_i) = [f']_i + \frac{[f'']_i}{2} h + O(h^2) - f'(x_i) = \frac{[f'']_i}{2} h + O(h^2) = O(h) = C_1 h,$$

где C_1 — некоторая константа, поскольку $f'(x_i) = [f']_i$. Если быть более точными, то для погрешности справедливо следующее

$$\Delta_{мет}^{(1)} = |f_i' - f'(x_i)| = O(h) = \max_{x \in [a,b]} |f''(x)| \frac{h}{2} = \frac{M_2 h}{2}. \quad (2.7)$$

Теперь рассмотрим метод, соответствующий схеме с центральной точкой, (2.4). Оценим погрешность этого метода. Необходимые разложения вокруг точки x_i

$$[f]_{i\pm1} = [f]_i \pm [f']_i h + \frac{[f'']_i}{2!} h^2 \pm \frac{[f''']_i}{3!} h^3 + O(h^4).$$

$$f_i' = \frac{[f]_{i+1} - [f]_{i-1}}{2h} = \frac{2[f']_i h + \frac{[f''']_i}{3} h^3 + O(h^3)}{2h} = [f']_i + \frac{[f''']_i}{6} h^2 + O(h^3).$$

Погрешность метода для (2.4) в таком случае равна

$$f_i' - f'(x_i) = [f']_i + \frac{[f''']_i}{6} h^2 + O(h^3) - f'(x_i) = \frac{[f''']_i}{6} h^2 + O(h^3) = O(h^2) = C_2 h^2,$$

где C_2 — некоторая константа. Таким образом,

$$\Delta_{мет}^{(2)} = |f_i' - f'(x_i)| = O(h^2) = \max_{x \in [a,b]} |f'''(x)| \frac{h^2}{6} = \frac{M_3 h^2}{6}. \quad (2.8)$$

Здесь и далее будет использоваться устоявшееся обозначение

$$M_n = \max_{\xi \in [a,b]} |f^{(n)}(\xi)|.$$

Как видим погрешность метода (2.7) на порядок по h меньше, чем (2.8). Таким образом, вводят понятие порядка (или порядка аппроксимации) метода. Более подробно о понят аппроксимация будет рассказано в параграфе, посвященном обыкновенным дифференциальным уравнениям, однако уже сейчас можно оперировать термином «порядок метода» по отношению к методам численного дифференцирования. Совершенно очевидно, что чем выше порядок метода, тем быстрей убывает погрешность с уменьшением шага.

Введем соответствующие определения погрешностей, которыми будем оперировать на протяжении курса.

Определение 2.1.

Пусть u и u^* — точное и приближенное значения некоторой величины соответсвенно. Тогда **абсолютной погрешностью приближения** u^* является $\Delta(u^*)$, удовлетворяющая соотношению

$$|u - u^*| \le \Delta(u^*).$$

Определение 2.2.

Относительной погрешностью приближения u^* является величина $\delta(u^*)$, удовлетворяющая соотношению

$$\left| \frac{u - u^*}{u^*} \right| \le \delta(u^*).$$

Определение 2.3.

Пусть искомая величина u является функцией, $u = u(t_1, t_2, ..., t_n)$, $t_1, t_2, ..., t_n \in \Omega$, u^* — приближенное значение u. Тогда **предельной абсолютной погрешностью функции** является величина

$$D(u^*) = \sup_{(t_1, t_2, ..., t_n) \in \Omega} \left| u(t_1, ..., t_n) - u^* \right|.$$

Определение 2.4.

Предельной относительной погрешностью функции $u = u(t_1, t_2, ..., t_n)$, $t_1, t_2, ..., t_n \in \Omega$ называется величина

$$d(u^*) = \frac{D(u^*)}{|u^*|}.$$

Пример 2.1. Погрешность функции

Пусть приближенное значение функции

$$y^* = y(a_1^*, a_2^*, ..., a_n^*),$$

рассмотрим наиболее распространенный случай, когда область Ω прямоугольная, тогда

$$|a_j - a_j^*| \le \Delta(a_j^*), \quad j = 1, ..., n.$$

Если функция y — непрерывно дифференцируемая, то применима формула Лагранжа для функции многих переменных,

$$y(a_1, a_2, ..., a_n) - y^* = \sum_{j=1}^{n} b_j(\theta)(a_j - a_j^*),$$

где $b_j(\theta) = \left. \frac{\partial y}{\partial a_j} \right|_{(a_1^* + \theta(a_1 - a_1^*), ..., a_n^* + \theta(a_n - a_n^*))}$, $0 \le \theta \le 1$. Отсюда справедлива следующая оценка погрешности

$$|y(a_1, a_2, ..., a_n) - y^*| \le D(y^*) = \sum_{j=1}^{n} B_j \Delta(a_j^*),$$

где $B_j = \sup_{\Omega} \left| \frac{\partial y}{\partial a_j} \right|$.

Положим $\rho = \sqrt{\sum_{j=1}^{n} \left(\Delta(a_j^*) \right)^2}$. Если все производные $\frac{\partial y}{\partial a_j}$ непрерывны, то $b_j(\theta) = b_j(0) + o(1)$, и в таком случае

$$B_j = \left| \frac{\partial y}{\partial a_j} \right\|_{a_1^*, ..., a_n^*} + o(1).$$

Тогда можно сделать следующую оценку

$$|y(a_1, a_2, ..., a_n) - y^*| \le D^*(y^*) = \sum_{j=1}^{n} \left| \frac{\partial y}{\partial a_j} \right\|_{a_1^*, ..., a_n^*} \Delta(a_j^*).$$

Сформулируем легко доказываемые свойства для погрешностей. Доказательство напрямую следует полученных выше соотношений.

Свойство 2.1.

Предельная погрешность суммы или разности равна сумме предельных погрешностей,

$$\Delta\left(\pm a_1^* \pm a_2^* \pm \ldots \pm a_n^*\right) = \Delta\left(a_1^*\right) + \Delta\left(a_2^*\right) + \ldots + \Delta\left(a_n^*\right).$$

Свойство 2.2.

Предельная относительная погрешность произведения или частного равна сумме предельных относительных погрешностей,

$$\delta\left(a_1^* \cdot a_2^* \cdot \ldots \cdot a_n^* \cdot b_1^{*-1} \cdot b_2^{*-1} \cdot \ldots \cdot b_m^{*-1}\right) = \delta\left(a_1^*\right) + \delta\left(a_2^*\right) + \ldots + \delta\left(a_n^*\right) + \delta\left(b_1^*\right) + \delta\left(b_2^*\right) + \ldots + \delta\left(b_m^*\right).$$

Пример 2.2. Пример оценки погрешности

Оценить погрешность приближения y^* к корню уравнения $a = f(y)$.

Значение правой части для приближенного решения

$$a^* = f\left(y^*\right).$$

При малом отклонении y^* от y соотношение

$$a^* - a = f\left(y^*\right) - f(y)$$

можно записать в виде

$$a^* - a = f\left(y^*\right) - f(y) \approx f'\left(y^*\right)\left(y^* - y\right).$$

Отсюда погрешность решения

$$y^* - y \approx \frac{a^* - a}{f'\left(y^*\right)} = \frac{f\left(y^*\right) - a}{f'\left(y^*\right)}.$$

В частном случае, когда $a = 0$, погрешность $y^* - y \approx \dfrac{f\left(y^*\right)}{f'\left(y^*\right)}$.

Пример 2.3. Погрешность неявной функции

Приведем оценку погрешности функции, заданной неявно, в общем виде.

$$F\left(y, a_1, \ldots, a_n\right) = 0.$$

Дифференцируя по a_j, имеем

$$\frac{\partial F}{\partial y}\frac{\partial y}{\partial a_j} + \frac{\partial F}{\partial a_j} = 0,$$

откуда

$$\frac{\partial y}{\partial a_j} = -\frac{\partial F}{\partial a_j}\left(\frac{\partial F}{\partial y}\right)^{-1}.$$

При заданных значениях a_j^*, $j = 1, \ldots, n$, приближенное значение функции можно найти как корень уравнения

$$F\left(y^*, a^*, \ldots, a_n^*\right) = 0.$$

Решая это уравнение, находим y^*, а затем значения выражений

$$\left.\frac{\partial y}{\partial a_j}\right|_{y^*, a_j^*, j=1,\ldots,n} = \left.-\frac{\partial F}{\partial a_j}\left(\frac{\partial F}{\partial y}\right)^{-1}\right|_{y^*, a_j^*, j=1,\ldots,n}.$$

Рассмотрим теперь способ представления чисел на ЭВМ. Основное ограничение хранения чисел — это разрядность числа. То есть, под число выделяется ограниченное

количество памяти или ограниченное число цифр. Одним из самых логичных форм хранения чисел является форма записи с плавающей точкой. Вещественное число a представляется в виде произведения мантиссы на основание системы счисления p в некоторой целой степени n, которую называют порядком:

$$a = m \cdot p^n .$$

Например, число 67.231 можно записать в следующем виде

$$67.231 = 0.67231 \cdot 10^2 ,$$

где $m = 0.67231$ — мантисса числа, $p = 10$ — основание десятеричной системы счисления, $n = 2$ — порядок. Порядок указывает, на какое количество позиций и в каком направлении должна «переплыть», то есть, сместиться десятичная точка в мантиссе. Отсюда название «плавающая точка».

Поясним эту форму записи на простом примере. Пусть для хранения числа 123.456789 имеется в распоряжении восемь десятеричных разрядов. Очевидно, что все число разместить в памяти не удается, поэтому производится отбрасывание последних цифр дробной части и округление. В первую очередь размещается целая часть числа. Если целая часть целиком не умещается в памяти, то производится отбрасывание старших цифр числа. В противном случае за целой частью размещается (плавающая) точка — разделитель целой и дробной частей. Далее все позиции заполняются оставшимися цифрами дробной части. На примере нашего числа пришлось отбросить последнюю цифру дробной части, произведя тем самым округление числа.

1	2	3	4	5	6	7	9

Число значащих цифр — число цифр дробной части, умещающиеся в разряды хранения числа. В нашем примере число имеет 5 значащих чисел. Любые арифметические операции над числами могут порождать дополнительные округления числа. При умножении числа, например, на 10 приводит к тому, что число значащих чисел числа сокращается с 5 до 4.

1	2	3	4	5	6	7	9

При умножении исходного числа, например, на 9 мы должны получить число 1111.11111. Однако при размещении этого числа в памяти сохранится лишь первые 8 цифр.

1	1	1	1	1	1	1	1	1

Таким образом, любая арифметика на ЭВМ — неизбежный процесс порождения округлений.

В связи с конечной разрядностью хранения чисел появляются числа, сложение с которыми никак не изменяет другие числа. Максимальное такое число называется машинным нулем.

Определение 2.5.

Машинным нулем называется максимальное число δ, такое, что

$$1 + \delta = 1$$

в ЭВМ-арифметике.

Каждый раз, когда вычисляется значение функции $f(x)$ в каком-либо сеточном узле x_i (см. (2.2)), появляется абсолютная погрешность $\delta_i = \delta \left| f \right|_i$. Действительно,

$$f_i \cdot 1 = f_i \cdot (1 + \delta) .$$

Исследуем методы первого (2.3) и второго (2.4) порядков на вычислительную погрешность. Погрешность округления для метода первого порядка равна

$$\Delta^{(1)}_{выч} = \frac{\delta_{i+1} + \delta_i}{h} .$$

Пусть, $\varepsilon = \max_i \delta_i$, тогда получаем, что

$$\Delta^{(1)}_{выч} = \frac{2\varepsilon}{h}.$$

Повторяя аналогичные рассуждения для метода второго порядка получаем.

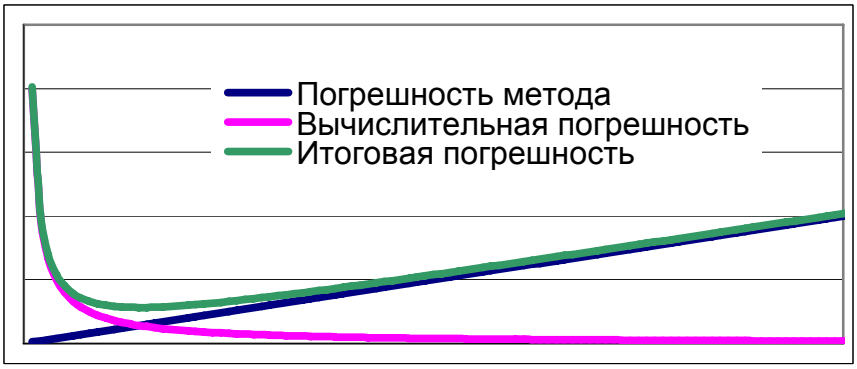
$$\Delta^{(2)}_{выч} = \frac{2\varepsilon}{2h} = \frac{\varepsilon}{h}.$$

Таким образом, теперь мы можем записать выражения для итоговых погрешностей этих двух методов соответственно

$$\Delta^{(1)} = \Delta^{(1)}_{мет} + \Delta^{(1)}_{выч} = \frac{M_2 h}{2} + \frac{2\varepsilon}{h},$$

$$\Delta^{(2)} = \Delta^{(2)}_{мет} + \Delta^{(2)}_{выч} = \frac{M_3 h^2}{3} + \frac{\varepsilon}{h}.$$

Качественно график зависимости итоговой погрешности от шага численного дифференцирования для каждого метода представлен на рисунке.



Из графиков видно, что существует некоторый оптимальный шаг дифференцирования, при котором итоговая погрешность минимальна, то есть, нет ярко выраженной погрешности метода или вычислительной погрешности. Этот шаг легко находится как экстремум функций $\Delta^{(1)}(h)$ и $\Delta^{(2)}(h)$. Отсюда оптимальные шаги для методов первого (2.3) и второго (2.4) порядков получаем

$$h^{(1)}_{опт} = 2\sqrt{\frac{\varepsilon}{M_2}}, \qquad h^{(2)}_{опт} = \sqrt[3]{\frac{3\varepsilon}{M_3}}.$$

Все слава, сказанные выше легко можно распространить на производные более высокого порядка.

Пример 2.4. Производная второго порядка

Получить формулу для вычисления производной второго порядка на неравномерной сетке.

Вторая производная по определению — первая производная от производной функции. Производные первого порядка в точках x_i и x_{i-1}

$$f'(x_{i-1}) \approx \frac{f_i - f_{i-1}}{h_{i-1}} \qquad \text{и} \qquad f'(x_i) \approx \frac{f_{i+1} - f_i}{h_i}.$$

соответственно. Тогда производную второго порядка можно записать следующим образом

$$f''(x_i) \approx \left(\frac{2}{h_i + h_{i-1}}\right)\left(f'(x_i) - f'(x_{i-1})\right) \approx \left(\frac{2}{h_i + h_{i-1}}\right)\left(\frac{f_{i+1} - f_i}{h_i} - \frac{f_i - f_{i-1}}{h_{i-1}}\right).$$

Для равномерной сетки с шагом h вторую производную можно вычислить по формуле

$$f''(x_i) \approx \left(\frac{2}{h+h}\right)\left(\frac{f_{i+1} - f_i}{h} - \frac{f_i - f_{i-1}}{h}\right) = \frac{f_{i+1} - 2f_i + f_{i-1}}{h^2}.$$