ЛЕКЦИЯ 1

ПОГРЕШНОСТЬ ВЫЧИСЛЕНИЙ

В наши дни, ни один крупный технический проект не обходится без различных расчетов и вычислений, начиная с очень простых алгебраических моделей, заканчивая сложнейшими научными расчетами, созданием алгоритмов, методов решения и т. д. В любом случае, необходимо знать на сколько полученный результат будет правильным.

Пусть a — истинное значение некоторой величины, a^* — ее приближенное значение.

Определение 1: *Абсолютная погрешность величины a^* называется наименьшее значение $\Delta(a^*)$, про которое известно, что*

$$|a^* - a| \leqslant \Delta(a^*).$$

Определение 2: *Относительная погрешность величины a^* называется наименьшее значение $\delta(a^*)$, про которое известно, что*

$$\left|\frac{a^* - a}{a^*}\right| \leqslant \delta(a^*).$$

Рассмотрим основные источники происхождения ошибок в вычислительной математике.

1. Неопределенность входных данных

Во многих задачах входными данными могут являться величины, значения которых были получены в ходе эксперимента. Однако, ни одну величину невозможно измерить абсолютно точно, всегда есть некоторый доверительный интервал, в котором лежит правильный результат.

Пример 1 x^* — полученное значение, $y^* \approx f(x^*)$ — значение функции.

$$x - \varepsilon \leqslant x^* \leqslant x + \varepsilon,$$

$$y - \delta \leqslant y^* \leqslant y + \delta.$$

Рассмотрим с какой точность можно определить значение функции. Пусть $f(x) = \sin x$.

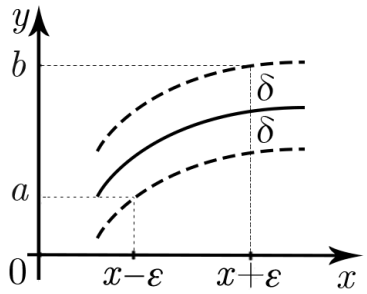

Рис. 1.1

Очевидно, что $|y^* - y| \leqslant |b - a|$. В данном примере такая оценка не самая лучшая. Можно выбрать

$$y^*_{опт} = \frac{b+a}{2}.$$

Тогда

$$|y^*_{опт} - y| \leqslant \frac{|b-a|}{2}.$$

Однако и такой выбор оптимального значения не всегда является наилучшим. *

2. Погрешность метода

Второй источник ошибок в вычислительной математике связан с погрешностью метода. Попробуем вычислить значение функции $y = f(x)$, где $f(x) = \sin x$. Для этого разложим ее в ряд Тейлора.

$$y = x - \frac{x^3}{3!} + \frac{x^5}{5!} + ...$$

С точки зрения вычислительной математики, в реальных вычислениях приходится ограничиваться конечным числом членов. В зависимости от того, сколько членов будет учтено, будет меняться погрешность метода.

$$y^*_1 = x + o(x^3),$$

$$y^*_2 = x - \frac{x^3}{3!} + o(x^5),$$

$$y^*_3 = x - \frac{x^3}{3!} + \frac{x^5}{5!}(x^3) + o(x^7).$$

Первый метод определяет значение функции с точностью до $o(x^3)$. Этот метод хорошо описывает погрешность вычисления для небольших x, но совершенно непременим, когда x достаточно большое. В данном примере

$$|y^*_n - y| = \frac{y^{(2n+1)}(\xi)}{(2n+1)!}x^{2n+1}, \quad \text{где } 0 < \xi < x.$$

3. Погрешность округления

Третий источник ошибок — это погрешность округления чисел. Вспомним представление машинных чисел. В современных компьютерах реализован стандарт IEEE арифметики. В IEEE арифметике действительные числа делятся на два класса: числа с одинарной точностью, для записи которых используются 32 бита (8 *байт*) и числа с двойной точностью, для записи которых используются 64 бита (16 *байт*).

Числа с одинарной точностью записываются следующим образом.

Знак	Степень	Мантисса
1	8	23

Один бит отводится под знак числа s, 8 бит под степень числа e и 23 бита под мантиссу числа f. Такие числа называются числами с плавающей запятой. Они представимы в следующем виде:

$$(-1)^s 2^{e-127}(1+f).$$

Это означает, что все числа на числовой оси расположены неравномерно и между каждой последовательной степенью двойки находится одно и то же количество модельных чисел. С таким представлением чисел связаны следующие важные константы вычислительной математики.

$\varepsilon_{маш}$ — машинная погрешность или погрешность округления.

Определение 3: *$\varepsilon_{маш}$ — наибольшая величина, для которой справедливо тождество*

$$1 + \varepsilon_{маш} = 1.$$

Исходя из представления числа видно, что $\varepsilon_{маш}$ равна половине последнего разряда. $\varepsilon_{маш} = 2^{-24} \approx 10^{-8}$. ♣

Определение 4: *OFL — Over Flow Limit (порог переполнения) максимальное число, которое может быть записано в арифметику.*
$OFL = 2^{256-127}(1+1) \approx 10^{38}$. ♣

Определение 5: *UFL — Under Flow Limit (порог машинного нуля) минимальное число, которое может быть записано в арифметику.*
$UFL = 2^{-127} \approx 10^{-38}$. ♣

Кроме нормализованных чисел в IEEE арифметике, где старший разряд равен 1, существуют субнормальные числа, где старший разряд может быть не только 1, но и 0. Они вводятся, для того чтобы правильно округлять результат математических операций, таких как сложение, вычитание, умножение и деление. Однако, может оказаться так, что результат операции над числами может не являться нормализованным числом. Обозначим как $\odot$ любую из этих операций $+, -, *, /$. Тогда операция округления $fl(a \odot b)$ округляет число до ближайшего числа с плавающей точкой. В этом случае округление считается правильным. Если результат лежит ровно между двумя числами, то операция округления округляет число до ближайшего четного с нулевым последним разрядом.

На числа с двойной точностью отводятся 64 бита (16 *байт*). Распределение между знаком числа, степенью и мантиссой будет следующим.

Знак	Степень	Мантисса
1	11	52

В этом случае число представимо в следующем виде

$$(-1)^s 2^{e-1023}(1+f).$$

Для чисел с двойной точностью получим:
$\varepsilon_{маш} = 2^{-53} \approx 10^{-16}, OFL = 2^{2048-1023} \approx 10^{308}, UFL = 2^{-1023} \approx 10^{-308}$.
В современных больших расчетах вычисления выполняются с двойной точностью.

Пример 2 Рассмотрим функцию $y = f(x)$, где

$$f(x) = \sum_{k=0}^{n} a_n x^k.$$

Для вычисления функции будем использовать схему Горнера. Она хороша тем, что является очень экономичным алгоритмом, при реализации схемы Горнера используется порядка $2n$ операций. Зададим схему следующим образом:

```
p=a_n
for k=n−1 to 0
p=px+a_n
end for
```

Пусть $f(x) = (x-2)^9 =$

$$= x^9 - 18x^8 + 144x^7 - 672x^6 + 2016x^5 - 4032x^4 + 5376x^3 - 4608x^2 + 2304x - 512.$$

Рассмотрим график функции. Точками обазначены значения, полученные используя схему Горнера.

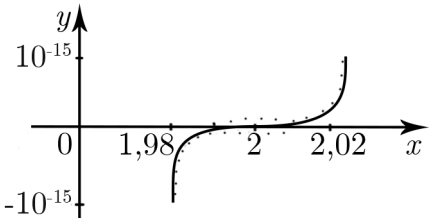

Рис. 1.2

При вычислении по схеме Горнера в окрестности точки $x = 2$ функция будет принимать значения различных знаков. При таком алгоритме невозможно определить положение нуля функции. Ошибка вызвана тем, что $y = f(x)$ содержит много положительных и отрицательных членов с большой степенью. Тем самым, вычитая и складывая числа порядка 2^9, мы пытаемся получить точность порядка 10^{-15}. Такое вычисление является некорректным. *

Пример 3 $f(x) = \mathrm{e}^x$.

$$\mathrm{e}^x = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + \ldots$$

Для вычисления $f(x)$ будем использовать тот факт, что начиная с некоторого номера, каждый следующий член ряда будет меньше чем ошибка округления. Рассмотрим алгоритм вычисления.

```
SUM=1
TERM=1
I=1
1.  TERM=TERM X/I
IF (SUM+TERM.EQ.SUM)
THEN
WRITE (*,*,X, SUM, EXP(X))
STOP
ELSE
SUM=SUM+TERM
GO TO 1
```

где `SUM` — сумма ряда, `TERM` — величина члена ряда.

Рассмотрим результат работы программы для чисел с одинаковой точностью.

X	SUM	EXP
1	2,718282	2,718282
5	148,4132	148,4132
10	22026,47	22026,46
15	3269017	3269017
20	$4{,}8516531 \cdot 10^8$	$4{,}8516521 \cdot 10^8$

X	SUM	EXP
-1	0,3678794	0,367795
-5	$6{,}7377836 \cdot 10^{-3}$	$6{,}7377947 \cdot 10^{-3}$
-10	$-1{,}6408609 \cdot 10^{-4}$	$4{,}539930 \cdot 10^{-4}$
-15	$-2{,}2377001 \cdot 10^{-7}$	$3{,}0590232 \cdot 10^{-7}$
-20	$1{,}202966 \cdot 10^{-9}$	$2{,}0611537 \cdot 10^{-9}$

Как можно видеть из приведенных таблиц, при достаточно больших положительных x ошибка остается маленькой, однако, при отрицательных значениях переменной функция меняет знак, что приводит к значительным ошибкам.

В этом примере основную ошибку дает алгоритм вычисления значения экспоненты при отрицательных x. Это связано с тем, что для отрицательных значений x ряд становится знакочередующимся. Нечетные степени образуют отрицательные величины. Для больших x наблюдается такой же эффект, как и в предыдущем примере. Значение экспоненты вычисляется, как сумма двух огромных значений. Так как количество разрядов ограничено, то на содержательное значение функции не остается численных разрядов. Такой алгоритм является неустойчивым, однако, его можно усовершенствовать.

Будем вычислять для $x < 0$ значения e^{-x}, а не e^x, как раньше. Тогда e^x вычислим, как

$$\mathrm{e}^x = \frac{1}{\mathrm{e}^{-x}}.$$

Такой алгоритм будет устойчивым. *

Пример 4 Рассмотрим систему уравнений следующего вида

$$\begin{cases} 0{,}780x + 0{,}563y = 0{,}217, \\ 0{,}457x + 0{,}330y = 0{,}127. \end{cases}$$

Нетрудно убедиться, что точным решением системы уравнений будут значения $x = 1{,}000$, $y = 1{,}000$. Если решать систему метод Гаусса с точностью три десятичных знака, то есть с точностью коэффициентов, то будут получены следующие значения $x = 1{,}71$, $y = -1{,}98$.

Подставим полученные значения в систему уравнений.

$$\begin{cases} 0{,}780(1{,}71) + 0{,}563(-1{,}98) - 0{,}217 = 0{,}00206, \\ 0{,}457(1{,}71) + 0{,}330(-1{,}98) - 0{,}127 = 0{,}00167. \end{cases}$$

Заметим, что, не смотря на довольно большие расхождения в значениях x и y, значения невязки оказались малыми. Это пример некорректно поставленной задачи (плохо обусловленной), когда небольшие изменения начальных входных данных приводят к сильным изменениям в ответе. *

Погрешность вычисления производной.

Во многих задачах математической физики неизвестен точный вид функции, известны только значения функции в некоторых точках, заданные в виде таблицы. В этом случае возникает необходимость вычислить производную по этим значениям. Вычисление производной можно приблизить различными способами. Например, используя разложение функции в трех точках: $f(x)$, $f(x+h)$, $f(x-h)$.

Используя ряд Тейлора получим

$$f(x+h) = f(x) + f'(x)h + \frac{1}{2}f''(x)h^2 + o(h^3),$$

$$f(x-h) = f(x) - f'(x)h + \frac{1}{2}f''(x)h^2 + o(h^3).$$

Исходя из этих соотношений, можно получить различные приближенные функции для вычисления производной.

1. **Производная вперед** $h > 0$

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

2. **Производная назад** $h < 0$

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

3. **Центральная разность**

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} + o(h^2).$$

Первые две формулы имеют погрешность поядка $o(h)$, они называются формулами первого порядка аппроксимации. Центральная разность имеет погрешность порядка $o(h^2)$, формула второго порядка аппроксимации.

Исходя из этих соотношений, можно сделать поспешный вывод, что чем меньше шаг h, тем более точный результат для вычисления производной будет получен. Однако это предположение ошибочно. Рассмотрим на примере формулы дифференцирования вперед.

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

Погрешность метода err состоит из двух ошибок. Из ошибки метода $\text{err}_{мет}$ и из ошибки округления $\text{err}_{окр}$. Для данного примера

$$\text{err}_{мет} = -\frac{1}{2}f''(\xi)h,$$

$$|\text{err}_{мет}| \leqslant \frac{1}{2}M_2 h,$$

где $M_2 = \max|f''(\xi)|$, $x \leqslant \xi \leqslant x+h$.

При малых h, существенный вклад в ошибку округления дает $\text{err}_{окр}$, связанная с тем, что значения $f(x)$ и $f(x+h)$ заданы с определенно точностью.

$$|\text{err}_{окр}| \leqslant \frac{M_0\varepsilon_{маш} + M_0\varepsilon_{маш}}{h},$$

где $M_0 = \max|f(\xi)|$, $x \leqslant \xi \leqslant x+h$.

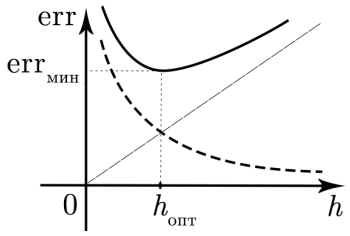

Рис. 1.3

Так как ошибка представляет собой сумму двух функций возрастающей и убывающей, то существует некоторое минимальное значение ошибки $\text{err}_{мин}$ и оптимальный шаг дифференцирования $h_{опт}$, при котором она достигается.

Вычислим оптимальный шаг дифференцирования.

$$\text{err} = \frac{2M_0\varepsilon_{маш}}{h} + \frac{1}{2}M_2 h,$$

$$\text{err}' = -\frac{2M_0\varepsilon_{маш}}{h^2} + \frac{1}{2}M_2 = 0 \Rightarrow h^2 = \frac{4M_0\varepsilon_{маш}}{M_2},$$

$$h_{опт} = 2\sqrt{\frac{M_0\varepsilon_{маш}}{M_2}}.$$

Тогда

$$\text{err}_{мин} = \frac{2M_0\varepsilon_{маш}}{2\sqrt{\frac{M_0\varepsilon_{маш}}{M_2}}} + \frac{1}{2}\cdot 2\sqrt{\frac{M_0\varepsilon_{маш}}{M_2}} = 2\sqrt{M_0 M_2 \varepsilon_{маш}}.$$

Пример 5 Пусть

$$f(x) = \sin x.$$

Тогда

$$M_0 = M_2 \approx 1,$$

$$h_{опт} = 4 \cdot 10^{-4}, \qquad \text{err} = 8 \cdot 10^{-4}.$$

Пусть теперь $f(x) = \sin 100x$.

$$M_0 = 1, \qquad M_2 = 10^4.$$

В этом случае оптимальный шаг уменьшится $h_{опт} = 4 \cdot 10^{-6}$, а ошибка увеличится $\text{err} = 8 \cdot 10^{-2}$.

Заметим, что при увеличении порядка аппроксимации, $h_{опт}$ тоже увеличится, а $\text{err}_{мин}$ уменьшится. Такие же оценки можно произвести для вычисления с двойной точностью, при $\varepsilon_{маш} = 10^{-16}$. *

Аналогичные формулы можно использовать и для вычисления второй производной. Вторая производная будет вычисляться, как производная от первой производной.

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

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

В этом случае погрешность метода

$$\text{err}_{мет} = \frac{1}{12} M_4 h^2,$$

где $M_4 = \max |f^{(4)}(\xi)|$, $x \leqslant \xi \leqslant x + h$.

$$\text{err}_{окр} \leqslant \frac{4M_0 \varepsilon_{маш}}{h^2}.$$

ЛИТЕРАТУРА

[1] *Калиткин Н.Н., Альшина Е.А.* Численный анализ. — М.: Академия, 2013.

[2] *Калиткин Н.Н., Корякин П.В.* Методы математической физики. — М.: Академия, 2013.

[3] *Вержбицкий В.М.* Численные методы. — М.: Высшая школа, 2000.

[4] *Лобанов А.И., Петров И.Б.* Численные методы.

[5] *Аристова Е.Н., Завъялова Н.А.,Лобанов А.И.* Практические занятия по вычислительной математике: учебное пособие. — М.: МФТИ, 2014.