ОСНОВЫ ТЕОРИИ ПОГРЕШНОСТЕЙ

Источники погрешности результата вычислений

- Математическая модель
- Исходные данные
- Приближённый метод
- Округления при вычислениях

Погрешности при численном решении задач делятся на две категории – **неустранимые и устранимые**.

Анализ погрешностей приближенных вычислений

- Погрешности задачи - приближённый характер математического описания поведения реального процесса вследствие неучёта или неправильного учёта существенных факторов, влияющих на результаты решения задачи (неустранимые или безусловные погрешности)
- Погрешности исходных данных –точность измерений экспериментальных данных (условно-устранимые)
- Обусловленность задачи – чувствительность её к малым изменениям входных параметров
- Погрешности методов решения задач – эффективность и надёжность метода (устранимые или условные)
- Погрешности алгоритма реализации численного метода решения задачи – сходимость к правильному решению после различного количества вычислений, влияющих на их суммарную погрешность:

- Переходная погрешность вычислительного процесса при пошаговой реализации выбранного алгоритма, передающаяся от итерации к итерации (накапливающаяся или самоисправляющаяся)
- Погрешность бесконечных вычислительных процессов – округление бесконечных рядов (остаточная погрешность)
- Погрешность округления из-за ограничения разрядности чисел, характерного для применяемого компьютера
- Погрешность действий – выполнение арифметических операций с приближенными числами

Все описанные типы погрешностей позволяют оценить полную погрешность результата решения прикладной задачи на компьютере.

Учёт погрешностей арифметических операций

Приближенным числом a называется число, незначительно отличающееся от точного A и заменяющее его в вычислениях.

Ошибка или погрешность Δa приближенного числа a – разность между точным и приближённым значениями: $\Delta a = A - a$

Абсолютная погрешность Δ приближенного числа a: $\Delta = |\Delta a|$

Относительная погрешность δ приближенного числа a: $\delta = \Delta / |A|$

Предельная абсолютная погрешность приближенного числа a – всякое число Δ_a, не меньшее абсолютной погрешности этого числа: $\Delta = |A - a| \leq \Delta_a$

Предельная относительная погрешность приближенного числа a – всякое число δ_a, не меньшее относительной погрешности этого числа: $\delta \leq \delta_a$

Числа Δ_a и δ_a называют *оценками* или *границами*, соответственно, абсолютной и относительной погрешностей.

Значащими цифрами числа в его позиционной записи называются все его цифры, начиная с первой ненулевой слева.

$a = 0{,}0\underline{2087}$ $a = 0{,}0\underline{208700}$ - значащие цифры подчёркнуты

Если приближённое число *a* имеет *n* значащих цифр, то за **предельную абсолютную погрешность** числа *a* принимают половину единицы разряда, выражаемого *n*–й значащей цифрой, считая слева направо.

Значащую цифру числа *a* называют **верной**, если абсолютная погрешность этого числа не превышает половины единицы разряда, соответствующего этой цифре (не превышает предельной абсолютной погрешности числа).

Вычислить приближённое число *a* с **точностью** до $\varepsilon=10^{-n}$ означает необходимость сохранить верной значащую цифру, стоящую в n-м разряде после запятой.

Пример Вычислить $\sqrt{2}$ с точностью $\varepsilon = 10^{-3}$.

Решение. $\sqrt{2} = \underline{1{,}414}2$; третья цифра после запятой является верной, т. к. $\Delta(\sqrt{2}) = |1{,}4142 - 1{,}414| = 0{,}0002 < 0{,}0005 = \Delta a < \varepsilon = 10^{-3}$. Следовательно, все подчеркнутые цифры являются верными.

Таким образом, из определения абсолютной погрешности приближенного числа a и точности его вычисления вытекает очевидная связь:

$$\Delta(a) \leqslant \Delta_a < \varepsilon,$$

т. е. *абсолютная погрешность и предельная абсолютная погрешность не превышают точности*.

Пример 1. Абсолютная и относительная погрешности приближенного числа *e*.

Число *e* - трансцендентное число, представляется бесконечной непериодической дробью $e = 2.71828$.

Приближенное значение числа $e^* = 2.7$, тогда
абсолютная погрешность $| e - e^* | = 0.01828$,
граница абсолютной погрешности 0.05,
относительная погрешность числа $| e - e^* | / e^* = 0.007$

Пример 2. Значащие цифры числа.

Значащие цифры чисел подчеркнуты:

$0.0\underline{3589}$, $\underline{10.4920}$, $0.00\underline{456200}$.

Пример 3. Верные цифры числа.

Верные цифры числа a = 356.78245 подчеркнуты.

- Если абсолютная погрешность числа $\Delta a = 0.01$, то верных цифр в числе 4:

  $a = \underline{356.7}8245$, т.к. при четырёх верных цифрах в этом числе его предельная абсолютная погрешность $\Delta_a = 0.05$, и соблюдается условие $\Delta a < \Delta_a$.
- Если $\Delta a = 0.03$, то верных цифр в числе также 4:

  $a = \underline{356.7}8245$ $\Delta_a = 0.05$, и соблюдается условие $\Delta a < \Delta_a$.
- Если $\Delta a = 0.00001$, то верных цифр в числе a будет 7:

  $a = \underline{356.7824}5$, при семи значащих цифрах $\Delta_a = 0.00005$, и $\Delta a < \Delta_a$.
- Если, $\Delta a = 0.00006$, то верных цифр в числе только 6:

  $a = \underline{356.782}45$ $\Delta_a = 0.0005$, и $\Delta a < \Delta_{a,}$, т.к. данная погрешность превышает предельную абсолютную погрешность, соответствующую 7 верным цифрам.

Поскольку истинные абсолютные и относительные погрешности неизвестны, то за них часто принимают предельные погрешности.

При округлении приближённого числа *a* до n-й значащей цифры необходимо к цифре (n+1)-го разряда прибавить цифру 5; если полученное число больше или равно 10, то к цифре n–го разряда добавляется единица, а разряды, начиная с (n+1)-го отбрасываются; в противном случае разряды начиная с (n+1)-го отбрасываются без прибавления единицы к n–му разряду.

Пример. Округлить число π= 3,141592:

а) до третьей значащей цифры;

б) до четвёртой значащей цифры.

Решение: а) 3,14, т.к. 1,592+5=6,592 < 10

б) 3,142, т.к. 5,92+5=10,92 > 10

Прямая задача теории погрешностей

Основная задача теории погрешностей состоит в том, чтобы определить по известным погрешностям параметров погрешность функции от этих параметров.

При грубом оценивании погрешности результата вычисления значения дифференцируемой функции $u=f(x_1, \dots, x_n)$ с приближёнными аргументами $x_1, \dots, x_n$ будем считать, что известны границы абсолютных погрешностей аргументов Δ_{x_1} , …, Δ_{xn} соответственно.

В этом случае точные значения аргументов $x^*_1, \dots, x^*_n$ лежат соответственно на отрезках:

$$\left[x_1 - \Delta x_1, x_1 + \Delta x_1\right], \dots, \left[x_n - \Delta x_n, x_n + \Delta x_n\right]$$

При этом абсолютная погрешность результата $u=f(x_1, \dots, x_n)$ равна

$$\Delta u = \left| f(x_1, \dots, x_n) - f(x_1^*, \dots x_n^*) \right|$$

и представляет собой модуль полного приращения функции.

Главной (линейной) частью этого приращения является полный дифференциал *du*, который имеет вид:

$$\Delta u \approx |du| = \left|\sum_{i=1}^{n} \frac{\partial u}{\partial x_1} dx_i\right| \leq \sum_{i=1}^{n} \left|\frac{\partial u}{\partial x_1}\right| \left|x_1 - x_1^{\star}\right| \leq \sum_{i=1}^{n} \left|\frac{\partial u}{\partial x_1}\right| \Delta_{x_i}$$

По этой причине за *границу абсолютной погрешности* результата приближенно может быть принята величина

$$\Delta_u = \sum_{i=1}^{n} \left|\frac{\partial u}{\partial x_i}\right| \Delta_{x_i} \quad (1.1)$$

В соответствии с этим равенством получается формула приближенной *оценки границы относительной погрешности* дифференцируемой функции *u*:

$$\delta_u = \frac{\Delta_u}{|u|} = \sum_{i=1}^{n} \left|\frac{\partial u}{\partial x_i}\right| \frac{\Delta_{x_i}}{|u|} = \sum_{i=1}^{n} \left|\frac{\partial u}{u \partial x_i}\right| \Delta_{x_i} = \sum_{i=1}^{n} \left|\frac{\partial \ln|u|}{\partial x_i}\right| \Delta_{x_i} \quad (1.2)$$

Оценка погрешностей арифметических действий

Оценка погрешностей арифметических действий проводится с применением формул (1.1) и (1.2).

- ***Сложение (вычитание)***

Пусть $u = x_1 + \ldots + x_n$. Тогда $\left|\frac{\partial u}{\partial x_i}\right| = 1$ и абсолютная погрешность операции сложения (вычитания) по формуле (1.1) будет равна

$$\Delta_{\sum(\pm x_i)} = \sum_{i=1}^{n} 1 \cdot \Delta_{x_i} = \sum_{i=1}^{n} \Delta_{x_i}$$

т.е. при сложении и вычитании приближенных чисел их предельные абсолютные погрешности складываются.

При оценке относительной погрешности суммы *положительных приближенных чисел* $x_1, x_2, \ldots, x_n$, имеющих границы

относительных погрешностей $\delta_{x_i}, \ldots, \delta_{x_n}$ соответственно, будет справедливо:

$$\delta(x_1 + \ldots + x_n) = \frac{\Delta(x_1 + \ldots + x_n)}{x_1 + \ldots + x_n} \le \frac{\Delta_{x_1 + \ldots x_n}}{x_1 + \ldots + x_n} = \frac{\Delta_{x_1} + \ldots + \Delta_{x_n}}{x_1 + \ldots + x_n} =$$

$$= \frac{x_1 \delta_{x_i} + \ldots + x_n \delta_{x_n}}{x_1 + \ldots + x_n} \le \frac{x_1 \delta^* + \ldots x_n \delta^*}{x_1 + \ldots + x_n} \le \delta^*$$

Где $\delta^* = \max \delta_{x_i} \ (1 \le i \le n)$

В результате можно сделать вывод, что относительная погрешность суммы положительных приближенных чисел не превосходит максимальной относительной погрешности слагаемых.

Для оценки относительной погрешности **разности** будет справедливо

$$\delta_{x_1 - x_2} = \frac{\Delta_{x_1 - x_2}}{|x_1 - x_2|} \le \frac{\Delta_{x_1} + \Delta_{x_2}}{|x_1 - x_2|}$$

Что указывает на возможность сильного возрастания погрешности при

$$x_1 - x_2 \to 0$$

В этом случае принято говорить о потере точности при вычитании близких чисел.

Умножение (деление)

В соответствии с формулой (1.2) при умножении положительных сомножителей $u = x_1 \cdot ... \cdot x_n$ будет справедливо

$$\ln u = \ln x_1 + ... + \ln x_n \qquad \frac{\partial \ln u}{\partial x_i} = \frac{1}{x_i}$$

В результате чего получается:

$$\delta_{\prod_{i=1}^{n} x_i} = \sum_{i=1}^{n} \frac{1}{x_i} \Delta_{x_i} = \sum_{i=1}^{n} \delta_{x_i}$$

При делении двух чисел $u = \frac{x_1}{x_2}$, где $x_1, x_2 > 0$

$$\ln u = \ln x_1 - \ln x_2 \qquad \left| \frac{\partial \ln u}{\partial x_i} \right| = \frac{1}{x_i}$$

И, соответственно, с учётом (1.2)

$$\delta_{x_1/x_2} = \frac{\Delta_{x_1}}{x_1} + \frac{\Delta_{x_2}}{x_2} = \delta_{x_1} + \delta_{x_2}$$

Таким образом, при умножении и делении приближенных чисел предельная относительная погрешность результата равна сумме предельных относительных погрешностей сомножителей.

Пример 4. Погрешности арифметических действий.

Пусть числа x и y заданы с абсолютными погрешностями Δx и Δy

$$x = 2.5378 \quad \Delta x = 0.0001 \qquad\qquad y = 2.536 \quad \Delta y = 0.001$$

Тогда относительные погрешности чисел будет равны:

$$\delta x = 3.94 \text{ x } 10^{-5} \qquad\qquad \delta y = 3.94 \text{ x } 10^{-4}$$

Найдем предельные абсолютные и относительные погрешности суммы и разности этих чисел.

$S1 = x + y \qquad \Delta\, S1 := \Delta_x + \Delta_y \qquad \delta_{S1}$

$S2 = x - y \qquad \Delta\, S2 := \Delta_x - \Delta_y \qquad \delta_{S2}$

....

Относительная погрешность разности более, чем в 1600 раз превышает относительную погрешность суммы!

3. Вычислить относительную погрешность в определении значения функции

$u = xy^2z^3$, если $x^* = 37,1, y^* = 9,87, z^* = 6,052, \Delta x^* = 0,3, \Delta y^* = 0,11, \Delta z^* = 0,016$.

Решение:

$$\delta_x = \frac{0.3}{37.1} \approx 0.81 \cdot 10^{-2}, \delta_y = \frac{0,11}{9,87} \approx 1,12 \cdot 10^{-2}, \delta_z = \frac{0,016}{6,052} \approx 0,26 \cdot 10^{-2},$$

$$\delta(u) = \delta(x^*) + 2\delta(y^*) + 3\delta(z^*) = 3.8 \cdot 10^{-2}.$$

Решение обратной задачи теории погрешностей

В отличие от прямой задачи оценивания погрешности результата вычисления значения функции при заданных оценках погрешностей аргументов обратная задача заключается в оценивании величин *Δxi* (и δx_i) по известной величине *Δu*.

С физической точки зрения необходимо определить, какой точности нужно подать данные на вход, чтобы получить результат заданной точности.

Для случая дифференцируемой функции одной переменной грубое решение обратной задачи тривиально: если $y = f(x)$

то $$\Delta y = |dy| = |f'(x)|\Delta x,$$

откуда $$\Delta x = \frac{\Delta y}{|f'(x)|}.$$

Для функции большого числа переменных нужно использовать дополнительные условия, формируемые, например, в соответствии с ***принципом равных влияний.***

Он состоит в предположении, что частные дифференциалы $\left|\frac{\partial u}{\partial x_i}\right|\Delta_{x_i}$ в (1.1) одинаково влияют на погрешность значения функции.

В результате получается, что в соответствии с (1.2)

$$\Delta_u = n\left|\frac{\partial u}{\partial x_i}\right|\Delta_{x_i}$$

Откуда решение обратной задачи имеет вид:

$$\Delta_{x_i} = \frac{\Delta_u}{n\left|\frac{\partial u}{\partial x_i}\right|}$$

Другим вариантом формулировки дополнительных условий является допущение о ***равенстве относительных погрешностей всех аргументов***, т.е. принимается, что

$$\delta_{x_i} = \frac{\Delta_{x_i}}{|x_i|} = \delta \qquad (i = 1, ..., n)$$

Тогда $\Delta_{x_i} = \delta |x_i|$ и в соответствии с (1.1) $\Delta_u = \delta \sum_{i=1}^{n} \left| \frac{\delta u}{\delta x_i} x_i \right|$

Из последнего равенства можно получить величину δ, характеризующую относительный уровень точности задания аргументов

$$\delta = \frac{\Delta_u}{\sum_{i=1}^{n} \left| x_i \frac{\partial u}{\delta x_i} \right|}$$

Откуда легко записываются выражения для определения границ абсолютных погрешностей аргументов:

$$\Delta_{x_i} = \frac{|x_i| \Delta_u}{\sum_{i=1}^{n} \left| x_i \frac{\partial u}{\partial x_i} \right|}$$

Статистический подход к учёту погрешностей арифметических действий

При больших количествах однотипных вычислений вступают в силу уже ***вероятностные*** или ***статистические законы*** формирования погрешностей результатов действий. Например, методами теории вероятностей показывается, что математическое ожидание абсолютной погрешности суммы n слагаемых с одинаковым уровнем абсолютных погрешностей, при достаточно большом n, пропорционально $\sqrt{n}$ ([20, 25, 61]). В частности, если $n > 10$ и все слагаемые округлены до m-го десятичного разряда, то для подсчета абсолютной погрешности суммы S применяют ***правило Чеботарева***

$$\Delta S \approx \sqrt{3n} \cdot 0.5 \cdot 10^{-m}. \qquad (1.4)$$

Пусть $x = \frac{1}{n}(x_1 + \ldots + x_n)$ — среднее арифметическое n ($n > 10$) приближенных чисел (например, результатов измерений), имеющих одинаковый уровень абсолютных погрешностей $\Delta_{x_i} = 0.5 \cdot 10^{-m}$. Тогда классическая оценка абсолютной погрешности величины x есть

$$\Delta_x = \frac{1}{n}\left(\Delta_{x_1} + \ldots + \Delta_{x_n}\right) = \frac{1}{n} \cdot n \cdot 0.5 \cdot 10^{-m} = 0.5 \cdot 10^{-m} = \Delta_{x_i},$$

т.е. такая же, как и у исходных данных. В то же время по формуле (1.4) имеем

$$\Delta_x \approx \frac{1}{n}\sqrt{3n} \cdot 0.5 \cdot 10^{-m} = \sqrt{\frac{3}{n}} \cdot 0.5 \cdot 10^{-m} = \sqrt{\frac{3}{n}} \cdot \Delta_{x_i} \xrightarrow[n \to \infty]{} 0.$$

Как видим, применение правила Чеботарева приводит к естественному выводу о том, что арифметическое усреднение результатов измерений или наблюдений увеличивает точность, чего нельзя сказать на основе классической теории погрешностей.

Технический подход к учёту погрешностей арифметических действий

Согласно принципу А.Н.Крылова, приближенное число должно записываться так, чтобы в нём все значащие цифры кроме последней были верными и лишь последняя была бы сомнительна, и при том в среднем не более, чем на единицу.

Чтобы результаты арифметических действий над приближенными числами, записанными таким образом, также соответствовали этому принципу, нужно придерживаться следующих правил:

1. *При сложении и вычитании приближенных чисел в результате следует сохранять столько десятичных знаков, сколько их в приближенном данном с наименьшим количеством десятичных знаков.*
2. *При умножении и делении в результате следует сохранять столько значащих цифр, сколько их имеет приближенное данное с наименьшим количеством значащих цифр.*
3. *Результаты промежуточных вычислений должны иметь один-два запасных знака (которые затем должны быть отброшены).*

Оценка погрешностей компьютерной арифметики

Основу памяти компьютера составляют базисные элементы, имеющие ***r*** устойчивых состояний (2, 8, 16 и т.п.). Каждому числу ставится в соответствие одинаковое количество ***k*** базисных элементов.

Упорядоченные базисные элементы образуют ***разрядную сетку машинного слова*** – в каждом разряде записано одно из базис-чисел 0, 1, 2,… ***r*** -1 и в специальном разряде - знак числа.

При записи числа с ***фиксированной запятой*** кроме ***r*** параметров основания системы счисления и ***k*** разрядов числа указывается ещё количество ***l*** разрядов под запись дробной части числа.

Таким образом, положительное вещественное число ***α*** может представлять собой в ***r***–ичной системе счисления дробь и отображается следующей конечной последовательностью:

$$a_1 a_2 \ldots a_{k-l} a_{k-l+1} \ldots a_{k-1} a_k$$

где $a_i \in \{0;1;\ldots;r-1\}$ т.е. реализуется приближенное равенство:

$$a \approx fix(a) = \alpha_1 r^{k-l-1} + \alpha_2 r^{k-l-2} + \ldots + \alpha_{k-l} r^0 + \alpha_{k-l+1} r^{-1} + \ldots + \alpha_{k-1} r^{-(l-1)} + \alpha_k r^{-l}$$

Абсолютная погрешность представления чисел с фиксированной запятой есть оценка величины $\left|a - fix(a)\right|$ от способа округления:

- r^{-1} *при простом отбрасывании* «хвоста» числа

$$\alpha_{k+1} r^{-(l+1)} + \alpha_{k+2} r^{-(l+2)} + \ldots$$

- *половина величины* r^{-1} *при округлении*, т.е. при увеличении α_k на единицу, если

$$\alpha_{k+1} > r/2$$

Абсолютная погрешность представления вещественных чисел с фиксированной запятой одинакова в любой части диапазона, в то время, как относительная погрешность

$$\left|\frac{a - fix(a)}{a}\right| \quad \text{или} \quad \left|\frac{a - fix(a)}{fix(a)}\right|$$

может значительно различаться в зависимости от того, берётся α близким к 0 или к границе диапазона .

Чаще употребляется представление вещественных чисел с плавающей запятой, которое записывается в экспоненциальной форме :

$$a = \pm M \cdot r^{p}$$

где ***r*** – основание системы счисления; ***p*** – порядок; ***M*** – мантисса числа, такая, что

$$r^{-1} \leq M < l(= r^{0})$$

Если под мантиссу выделяется ***l*** ***r***-ичных элементов, а под порядок – ***m***,

то в системе записи с плавающей запятой вещественное число ***a*** представляется конечным числом ***fl(a)***

$$a = fl(a) = \pm(\beta_1 r^{-1} + \beta_2 r^{-2} + \ldots + \beta_l r^{-l}) r^{\gamma}$$

Где **r**– целое число из промежутка

$$\left[-r^{m}, r^{m} - 1\right] \qquad \beta_1 \in \left[1;\ldots;r-1\right] \qquad \beta_i \in \left[0;1;\ldots;r-1\right] \qquad (i = 2,\ldots l)$$

т.е. машинное слово имеет следующую структуру:

Знак порядка	Порядок	Знак мантиссы	Мантисса

m разрядов

l разрядов

Числа $\pm r^{r^m}$ определяют границы допустимого числового диапазона, при этом диапазон представления положительных вещественных чисел составляет промежуток

$$[r^{-r^m}, r^{r^m-1}]$$

Левую и правую границы этого отрезка называют соответственно ***машинным нулём и машинной бесконечностью***, так как числа из промежутка $[-r^{-r^m}, r^{r^m-1}]$ компьютер заменяет нулём, а числа, лежащие за пределами промежутка $[-r^{-r^m-1}, r^{r^m-1}]$ он не воспринимает.

Важной характеристикой является число ε, называемое ***машинным эпсилон*** и обозначаемое обычно идентификатором ***macheps***. Определяется как расстояние между единицей и ближайшим следующим за ней числом системы машинных чисел с плавающей запятой. Так как

$$1 = \left(1r^{-1} + 0r^{-2} + \ldots + 0r^{-l} + \ldots\right) \cdot r^{1}$$

а следующее за 1 машинное число есть

$$\left(1r^{-1} + 0r^{-2} + \ldots + 0r^{-(l-1)} + 1r^{-l}\right) r^{1} = fl(1+\varepsilon)$$

то за ***macheps*** можно принять величину

$$\varepsilon = 1r^{-l} r^{1} = r^{1-l}$$

Это число непосредственно связано с относительной погрешностью представления чисел в системе с плавающей запятой

$$\left|\frac{a - fl(a)}{a}\right| = \frac{\beta_{l+1} r^{-(l+1)} + \beta_{l+2} r^{-(l+2)} + \ldots}{\beta_1 r^{-1} + \beta_2 r^{-2} + \ldots} \le \frac{1 \cdot r^{-l}}{\beta_1 \cdot r^{-1}} \le r^{(1-l)} = \varepsilon$$

Таким образом, ***машинный эпсилон*** служит мерой относительной погрешности представления вещественных чисел, причём эта точность одинакова в любой части числового диапазона и зависит лишь от числа ***r***-ичных разрядов, отводимых под мантиссу числа.

В то же время оценка абсолютной погрешности

$$\left|a - fl(a)\right| \le |a| \cdot r^{(1-l)}$$

показывает, что расстояние между вещественными числами и конечными приближениями к ним в системе с плавающей запятой не одинаковы в разных частях числового диапазона.

Величина ***macheps*** служит оценкой относительной точности представления вещественного числа ***α*** при условии, что

$$|a| > r^{-r^m}$$

Если $a \in [-r^{-r^m}, r^{-r^m}]$, то $fl(a) \equiv 0$ и это значит, относительная

погрешность $\left|\dfrac{a - fl(a)}{a}\right| \equiv 1$

Т.е. является постоянной достаточно большой величиной, в то время

как абсолютная погрешность не превосходит величины r^{-r^m}

Пример.

Для записи числа в 48-разрядном машинном слове 40 двоичных разрядов выделяется под мантиссу, 6 – под порядок и 2 – под знаки мантиссы и порядка. Отсюда, принимая, что ***r*** = 2, ***l*** = 40, ***m*** = 6, получаем, что точность представления чисел с плавающей запятой не хуже 2^{-39} ($\sim 10^{-12}$), граница ***машинного нуля*** 2^{-64} ($\sim 10^{-19}$), ***машинной бесконечности*** 2^{63} ($\sim 10^{19}$).

Если машинное слово имеет 32 двоичных разряда, из них под мантиссу выделяется 24, а под порядок 7. Зная параметры ***r*** = 2,

l = 24, ***m*** = 7, получаем ***macheps*** = 2^{-23} ($\sim 10^{-7}$), ***машинные нуль*** 10^{-38} и ***бесконечность*** = 10^{38}

Когда используется представление вещественных чисел по основанию ***r*** = 16, эти машины имеют относительную точность представления $\sim 10^{-7}$ и диапазон для положительных чисел $\sim 10^{-77} \div 10^{76}$.

Практически в любом компьютере промежуточные вычисления производятся с двойной точностью, что учитывается также и большинством языков программирования.

Возникновение возможных больших случайных погрешностей обусловлено следующими основными причинами:

- методом округления, принятом в компьютере;
- потерями значащих разрядов при вычитании;
- потерей разрядов при превышении допустимой разрядности представления чисел (например при делении на малые числа).

Устойчивость численного метода

Под ***устойчивостью численного (приближенного)*** метода подразумевается несущественное отклонение получаемых приближенных результатов от точного решения.

Строго:

Численный метод называется устойчивым, если для любой погрешности ***ε>0*** в исходных данных существует такое ***δ>0***, что максимальная погрешность результатов будет меньше ***ε*** при максимальной погрешности ввода, меньшей ***δ.***

Основной задачей при реализации численных методов (комбинации численных методов) для решения задач компьютерного моделирования является обеспечение их устойчивости, т.е. минимизации всевозможных погрешностей.

Обусловленность задачи

Пример 1.1. Вычислить все корни уравнения

$$x^4 - 4x^3 + 8x^2 - 16x + 15.\underbrace{99999999}_{8} = (x-2)^4 - 10^{-8} = 0.$$

Точное решение задачи легко найти:

$$(x-2)^2 = \pm 10^{-4},$$

$$x_1 = 2,01; \quad x_2 = 1,99; \quad x_{3,4} = 2 \pm 0,01i.$$

Если компьютер работает при $\delta_M > 10^{-8}$, то свободный член в исходном уравнении будет округлен до 16,0 и, с точки зрения представления чисел с плавающей точкой, будет решаться уравнение $(x-2)^4 = 0$, т. е. $x_{1,2,3,4} = 2$, что, очевидно, неверно. В данном случае малые погрешности в задании свободного члена $\approx 10^{-8}$ привели, независимо от метода решения, к погрешности в решении $\approx 10^{-2}$.

Пример 1.4. Решением системы линейных алгебраических уравнений (СЛАУ)

$$\begin{cases} u + 10v = 11 \\ 100u + 1001v = 1101 \end{cases}$$

является пара чисел $\{1, 1\}$.

Изменив правую часть системы на 0,01, получим возмущенную систему

$$\begin{cases} u + 10v = 11.01 \\ 100u + 1001v = 1101 \end{cases}$$

с решением $\{11.01; 0.00\}$, сильно отличающимся от решения невозмущенной системы.Эта система также плохо обусловлена.

Влияние выбора вычислительного алгоритма на результаты вычислений

Пример 1.6. Пусть необходимо вычислить значение выражения $\left(\frac{\sqrt{2}-1}{\sqrt{2}+1}\right)^3$.

Избавившись от знаменателя, получаем $(\sqrt{2}-1)^6 = (3-2\sqrt{2})^3 = 99-70\sqrt{2}$.

Полагая а) $\sqrt{2} \approx \frac{7}{5} = 1,4$, в) $\sqrt{2} \approx \frac{17}{12} = 1,41(6)$ и рассматривая эти приближения как разные методы вычисления, получим следующие результаты:

$\sqrt{2}$	$(\sqrt{2}-1)^6$	$(3-2\sqrt{2})^3$	$99-70\sqrt{3}$
7/5	0,004096	0,008000	1
17/12	0,005233	0,004630	-0,1(6)

Очевидно, что столь значительное различие в результатах вызвано влиянием ошибки округления в задании $\sqrt{2}$.

Экономичность вычислительного метода

Пример 1.12. Пусть требуется вычислить сумму $S = 1 + x + x^2 + \ldots + x^{1023}$ при $0 < x < 1$. Для последовательного вычисления $x, x^2 = x \cdot x, \ldots, x^{1023} = x^{1022} \cdot x$ необходимо проделать 1022 умножения, а затем столько же сложений.

Однако если заметить, что $S = \frac{1-x^{1024}}{1-x}$, то количество арифметических действий значительно уменьшается; в частности, для вычисления x^{1024} требуется всего 10 умножений: $x^2 = x \cdot x; x^4 = (x)^2(x)^2, \ldots, x^{1024} = (x)^{512}(x)^{512}$.

Задание.

1. Самим выполнить пример 4 из лекции.

2. Вычислить абсолютную и относительную погрешности функции многих переменных $u(x,y,z)=x^2y^2/z^4$, если заданы $x=37.1$ $y=9.87$ $z=6.052$ $\Delta x = 0.1$ $\Delta y = 0.05$ $\Delta z = 0.02$

3. Вычислить абсолютную и относительную погрешности функции многих переменных.

Пусть $x = -3.59$ $y = 0.467$ $z = 563.2$ По приведенным начальным условиям считаем, что погрешности переменных равны $\Delta x = 0.01$ $\Delta y = 0.001$ $\Delta z = 0.1$