Теория погрешностей и машинная арифметика

Калитвин В.А.
kalitvinv@yandex.ru

2025

Способы представления чисел

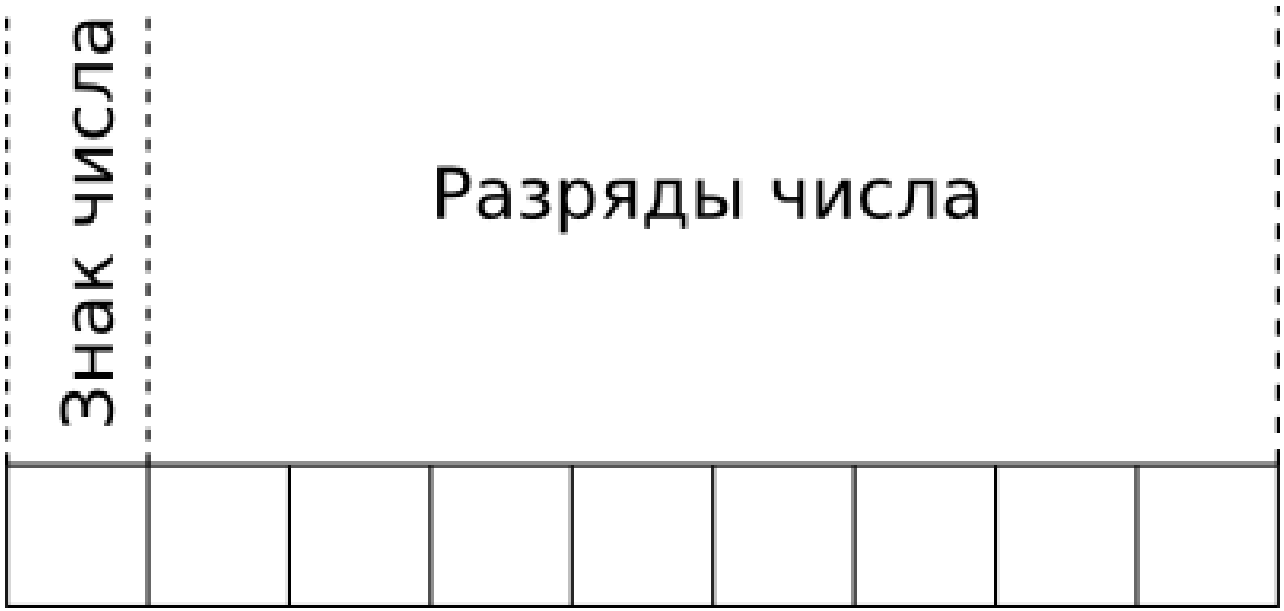
В ЭВМ используются в основном два способа представления чисел: *с естественным размещением запятой* и в форме с так называемой *плавающей запятой.*

При использовании формы с естественным размещением, запятая может располагаться в любом цифровом разряде сетки, а ее место определяется или при вводе числа с клавиатуры, или в результате выполнения операции.

Легко видеть, что в МК с 8-разрядной сеткой в форме с естественным размещением запятой могут быть представлены лишь числа от $\pm 10^{-7}$ до $\pm(10^8 - 1)$.

Способы представления чисел

Естественная форма представления чисел



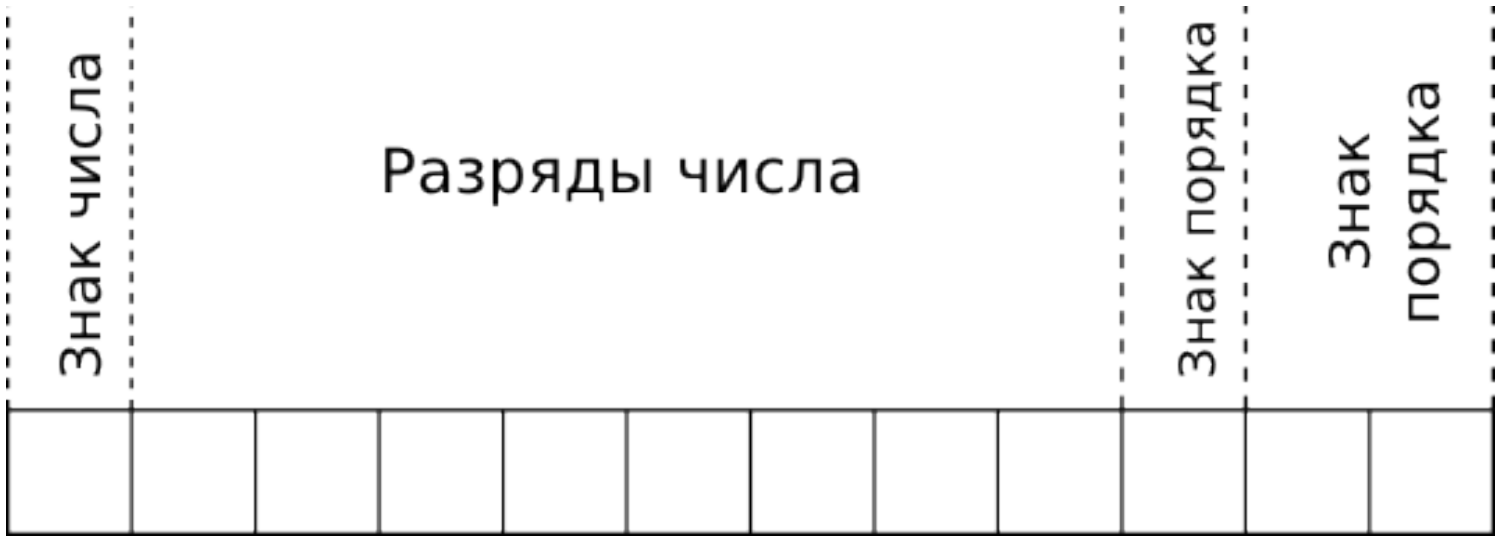
Способы представления чисел

Способ представления чисел в форме с плавающей запятой имеет вид

$$a = M \cdot 10^p.$$

Число M называется мантиссой, p - показателем.

Форма с плавающей запятой



Способы представления чисел

Чтобы избежать неоднозначности представления чисел, испольуют *нормализованные числа.*

Например, если $0,1 \leq M < 1$, то число a называют нормализованным.

У обоих чисел M и p количество разрядов конечно, поэтому имеется лишь конечное множество чисел с плавающей точкой.

Следовательно, сущестуют наибольшее и наименьшее числа с плавающей точкой, т.е. все представимые на ЭВМ числа удовлетворяют неравенству:

$$0 < X_0 \leq |x| < X_\infty.$$

Способы представления чисел

Если результат операции превышает X_∞, то происходит переполнение, и для большинства компьютеров данное вычисление на этом заканчивается. Все числа по модулю большие X_∞, не представимы на ЭВМ и рассматриваются как *машинная бесконечность*.
Если результатом операции является число, слишком близкое к нулю, то компьютеры заменяют результат нулем, т.е. все числа по модулю меньшие X_0, рассматриваются как *машинный нуль*.
При сложении машинных чисел различной величины результат может оказаться точно равен одному из слагаемых. Наименьшее число с плавающей запятой, которое при сложении с числом 1 дает результат, больший чем 1, называется *машинным эпсилон* и обозначается ε_M. Машинный эпсилон определяет относительную погрешность арифметики компьютера.

Способы представления чисел

Если x и y — два положительных числа с плавающей точкой и $x > y$, то их сумму можно записать в виде

$$x + y = x\left(1 + \frac{y}{x}\right).$$

Очевидно, что при $\frac{y}{x} < \varepsilon_M$ сумма с плавающей точкой чисел x и y совпадает с x. Более тщательное исследование показывает, что относительная погрешность сложения чисел с плавающей точкой ограничена величиной ε_M.

Числа с плавающей точкой между 0 и наибольшим числом с плавающей запятой распределены неравномерно.

Между каждыми двумя соседними степенями двойки находится 2^{22} чисел с плавающей точкой, например 2^{22} чисел между 2^{-128} и 2^{-127} и столько же чисел между 2^{126} и 2^{127}. Таким образом, числа с плавающей точкой гуще расположены вблизи нуля.

Стандарт IEEE 754

IEEE (Institute of Electrical and Electronics Engineers) (Институт инженеров по электротехнике и электронике) — международная ассоциация специалистов в области техники, мировой лидер в области разработки стандартов по радиоэлектронике и электротехнике.

IEEE 754 — широко распространённый стандарт формата представления чисел с плавающей точкой в двоичном коде. Используется многими аппартными средствами (CPU и FPU) и программными реализациями. Многие компиляторы языков программирования используют этот стандарт для хранения данных и выполнения математических операций.

Стандарт IEEE 754

Стандарт описывает:
Определения форматов хранения мантиссы, показателя и знака, форматы положительного и отрицательного нуля, плюс и минус бесконечностей, а также определение «не числа» (NaN).
Округления (округление к ближайшему, прямое округление, точность округления).
Операции (арифметика, квадратный корень, методы, которые используются для преобразования форматов чисел с плавающей точкой, преобразования между форматами с плавающими точками и форматами целых чисел, округление чисел с плавающей точкой в целые числа, преобразование бинарного представления в десятичное, сравнение.
Обработку исключительных ситуаций, таких как деление на нуль, переполнение, потеря значимости,
Обнаружение недопустимых операций.

Стандарт IEEE 754

В компьютере с двоичной арифметикой числа с плавающей точкой обычно представляются в памяти 32 двоичными разрядами, или битами. Стандарт IEEE отводит 24 бита для мантиссы и 8 битов для показателя. Сюда также входят биты для хранения знаков мантиссы и показателя. Предполагается, что самый левый бит мантиссы любого числа равен 1, что позволяет не хранить этот бит. Наибольшее число с плавающей точкой $\approx 10^{38}$. Наименьшее число с плавающей точкой $\approx 10^{-38}$. Мантисса из 24 битов соответствует примерно 7 десятичным разрядам.

Стандарт IEEE 754

Для 32-битовой арифметики с плавающей точкой, удовлетворяющей стандарту IEEE, $\varepsilon_M = 2^{-32} \approx 1.2 \times 10^{-7}$ при использовании округлений. Поэтому бессмысленно рассчитывать более чем на семь верных десятичных знаков в любом результате вычисления с плавающей точкой или на то, что мы сумеем разрешить относительные различия, меньшие этого уровня.

Стандарт IEEE рекомендует также, чтобы компьютеры выполняли арифметику с большей разрядностью, несмотря на то, что результаты операций записываются в память с 32 битами. В устройствах, поддерживающих этот стандарт, арифметика с плавающей точкой нередко реализована с внутренней разрядностью 80 битов.

Стандарт IEEE 754

Стандарт IEEE 754 широко применяется в технике и программировании. Большинство современных микропроцессоров изготовляются с аппаратной реализацией представления вещественных переменных в формате IEEE 754. Язык программирования и программист не могут изменить эту ситуацию, иного преставления вещественного числа в микропроцессоре не существует. Когда создавали стандарт IEEE754-1985 представление вещественной переменной в виде 4 или 8 байт казалось очень большой величиной, так как объём оперативной памяти MS-DOS был равен 1 Мб. А, программа в этой системе могла использовать только 0,64 Мб. Для современных ОС размер в 8 байт является ничтожным, тем не менее переменные в большинстве микропроцессоров продолжают представлять в формате IEEE 754-1985.

Основные понятия теории прибл. вычислений

Пусть X — истинное значение некоторой величины, а x – ее известное приближение.
Абсолютной погрешностью приближенного значения x называется величина

$$e_x = |X - x|.$$

Величина e_x в большинстве случаев остается неизвестной для вычислителя, так как для ее вычисления нужно знать точное значение X.

На практике обычно удается установить верхнюю границу абсолютной погрешности, т.е. такое (по возможности наименьшее) число Δx, для которого справедливо

$$|X - x| \leq \Delta X.$$

Число Δx в этом случае называют **границей абсолютной погрешности** (или *предельной абсолютной погрешностью)* приближения x.
Неравенство $|X - x| \leq \Delta x$ позволяет установить приближения к точному значению X по недостатку и избытку:

$$|x - \Delta x| \leq X \leq |x + \Delta x|.$$

Качество приближенных значений измеряется с помощью относительной погрешности, которая определяется как отношение ошибки e_x к модулю значения X (когда оно неизвестно — к модулю приближения x).
Границей относительной погрешности δx приближенного числа называется отношение предельной абсолютной погрешности к абсолютному значению приближения x :

$$\delta x = \frac{\Delta x}{|x|}.$$

Относительную погрешность обычно выражают в процентах.

Цифра числа называется *верной (в широком смысле)*, если абсолютная погрешность числа не превосходит единицы разряда, в котором стоит эта цифра.
Значащими цифрами числа, записанного в виде десятичной дроби, называют все его верные цифры, начиная с первой слева, отличной от нуля.
Запись приближенного числа только верными знаками принято называть *правильной записью.*
Исключение из последующих вычислений неверных цифр производится путем округления приближенных чисел, т.е. замены числа его значением с меньшим количеством значащих цифр.

При округлении возникает погрешность, называемая *погрешностью округления*. Пусть x — данное число, а x_1 — результат его округления. Погрешность округления определяется как модуль разности прежнего и нового значений числа:

$$\Delta_{\text{окр.}} = |x - x_1|$$

В отдельных случаях вместо $\Delta_{\text{окр.}}$ приходится использовать ее верхнюю оценку.

В основном испольются округление методом отбрасывания и симметричного округления.
При округлении методом отбрасывания цифры, стоящие после разряда, до которого производится округление, отбрасываются. Сам по себе метод отбрасывания оставляет все сохраняемые цифры округленного числа верными.
Способ симметричного округления приводит к меньшей величине ошибки округления, чем способ отбрасывания. Симметричное округление выполняется по правилам:
1.Если первая слева, из всех отбрасываемых цифр меньше 5, то последняя сохраняемая цифра остается без изменения.
2. Если первая слева, из всех отбрасываемых цифр больше или равна 5, то последняя сохраняемая цифра увеличивается на единицу.

Из правил симметричного округления следует, что его погрешность не превышает половины единицы последнего сохраняемого разряда. Это обстоятельство позволяет вести учет с точностью большей, чем единица последнего сохраняемого разряда. По этой причине наряду с понятием *«верная цифра»*, соответствующим методике округления путем отбрасывания, используется понятие *«цифра, верная в строгом смысле»*, применяемое в вычислениях с симметричным округлением.

Цифра числа называется *верной в строгом смысле*, если абсолютная погрешность этого числа не превосходит половины единицы разряда, в котором стоит эта цифра.
Абсолютная погрешность числа x_1, получаемого в результате округления приближенного значения x, складывается из абсолютной погрешности первоначального числа x и погрешности округления. Действительно, из неравенства

$$|X - x_1| \leq |X - x| + |x - x_1| \leq \Delta x + \Delta_{\text{окр}}.$$

следует, что если в результате округления приближенного числа получено значение x_1, то границей абсолютной погрешности числа x_1 можно считать сумму границы абсолютной погрешности числа x и погрешности округления.

Определение количества верных цифр по относительной погрешности

Количество верных значащих цифр в приближенном числе и величина относительной погрешности этого числа взаимосвязаны: по величине δx можно вычислять абсолютную погрешность

$$\Delta x = |x| \cdot \delta x,$$

величина которой влияет на количество верных значащих цифр в приближенном числе. На практике иногда удобнее пользоваться правилом, устанавливающим взаимосвязь количества верных цифр непосредственно с величиной относительной погрешности.

Определение количества верных цифр по относительной погрешности

Рассмотрим приближенное число x, записанное в нормальном виде

$$x = M \cdot 10^p, 0,1 \leq M < 1 \tag{1}$$

(число x называют в этом случае нормализованным).
Заметим, что для нормализованного числа x справедливо

$$|x| < 10^p.$$

Определение количества верных цифр по относительной погрешности

Итак, имеется приближенное число x и его относительная погрешность δx. Нужно установить количество верных в строгом смысле значащих цифр в числе x.
Для каждого известного значения δx можно подобрать такое наибольшее натуральное n, чтобы имело место

$$\delta x \leq 10^{-n}.$$

Тогда

$$\Delta x \leq |x| \cdot 10^{-n} < 10^{p} \cdot 10^{-n} = 10^{p-n} < \frac{1}{2} 10^{-(n-1)+p},$$

$$\Delta x < \frac{1}{2} 10^{-(n-1)} \cdot 10^{p}. \tag{2}$$

Определение количества верных цифр по относительной погрешности

Сопоставляя теперь (1) и (2) и используя определение цифры, верной в строгом смысле, можно сделать вывод, что в мантиссе приближенного числа x верны в строгом смысле, по крайней мере $n-1$ цифра после запятой (а поскольку x нормализовано, то все эти цифры значащие).

Следовательно

Для того чтобы по заданной величине δx найти количество верных значащих цифр в числе x, достаточно подобрать наибольшее натуральное число n так, чтобы выполнялось неравенство $\delta x \leq 10^{-n}$, и уменьшить n на единицу.

Определение количества верных цифр по относительной погрешности

Пример. Пусть $x = 15,327, \quad \delta x = 0,007.$
Так как $0,0007 \leq 10^{-3}$, то x имеет не менее двух верных в строгом смысле цифр.
В некоторых случаях в числе x могут оказаться верными n цифр.
Если первая значащая цифра в значении δx меньше 5, то можно использовать условие

$$\delta x \leq \frac{1}{2} \cdot 10^{-n}.$$

Тогда число x имеет не менее n верных в строгом смысле цифр.

Определение количества верных цифр по относительной погрешности

Действительно,

$$\Delta x = |x| \cdot \delta x \leq \frac{1}{2}|x| \cdot 10^{-n} \leq \frac{1}{2}10^{p} \cdot 10^{-n},$$

а это означает, что мантисса M нормализованного числа x имеет по меньшей мере n верных в строгом смысле цифр.

Пример. Пусть $x = 127,453,\ \delta x = 0,004.$

Так как $0,004 \leq 0,005 = \frac{1}{2} \cdot 10^{-2}$, то в x являются верными в строгом смысле не менее двух цифр.

Сумма и разность

Пусть $S = X + Y$ — сумма точных чисел, а $s = x + y$ — сумма их приближеных значений.
Тогда

$$\Delta s = \Delta x + \Delta y.$$

$$S - s = (X - x) + (Y - y).$$

$$|S - s| \leq |X - x| + |Y - y|,$$

т.е. $e_s \leq e_x + e_y$, или $e_s \leq \Delta x + \Delta y$. Следовательно, можно принять

$$\Delta s = \Delta x + \Delta y$$

Пример. Даны приближенные значения $x = 235,4$ и $y = 79,1834$, у которых все цифры являются верными в широком смысле. Найдем их сумму

$$S = 235,4 + 79,1834 = 314,5834.$$

Для оценки точности результата вычислим сумму погрешностей слагаемых $10^{-1} + 10^{-4} = 0,1001$. Величина ошибки показывает, что в результате уже первый знак после запятой является сомнительным.

Из приведенного примера следует:
1. получение результата с большим числом цифр еще не означает, что все эти цифры верны.
2. при вычислении сумм и разностей чисел с сильно отличающимися абсолютными погрешностями целесообразно округлить исходные данные, оставив столько десятичных знаков, сколько их имеет слагаемое с наименьшим числом десятичных знаков.
Заметим, что при последовательном вычитании и сложении нескольких чисел выгоднее производить действия над числами в порядке возрастания их абсолютных величин.

Пример.
Пусть требуется найти сумму пяти четырехразрядных чисел:

$$S = 0.2764 + 0.3944 + 1.475 + 26.46 + 1364$$

Складывая все эти числа, а затем, округляя результат до четырех значащих цифр, получаем $S = 1393$.
Однако при вычислениях на компьютере округление происходит после каждой операции сложения.
Предполагая условно сетку четырех разрядной ($k = 4$), проследим за вычислениями на компьютере суммы чисел в порядке их записи

0.2764+0.3944=0.6708, 0.6708+1.475=2.156,
2.156+26.46=28.62, 28.62+1364=1393;
получили $S_1 = 1393$, т.е. верный результат.
Изменим теперь порядок вычислений и начнем складывать числа последовательно от последнего к первому:
1364+26.46=1390, 1390+1.475=1391,
1391+0.3944=1391, 1391+0.2764=1391;
здесь окончательный результат $S_2 = 1391$, он менее точный.

Анализ процесса вычислений показывает, что потеря точности здесь происходит из-за того, что прибавления к большому числу малых чисел не происходит, поскольку они выходят за рамки разрядной сетки. Таких малых чисел может быть очень много, но на результат они все равно не повлияют, поскольку прибавляются по одному, что и имело место при вычислении S_2. Здесь необходимо придерживаться правила, в соответствии с которым сложение чисел следует проводить по мере их возрастания. Таким образом, в машинной арифметике из-за погрешностей округлений существенен порядок выполнения операций.

Относительные погрешности суммы и разности можно вычислить через абсолютные, но можно использовать и специальные формулы

$$\delta\left(x+y\right)=\frac{|x|}{|x+y|}\cdot\delta x+\frac{|y|}{|x+y|}\cdot\delta y,$$

$$\delta\left(x-y\right)=\frac{|x|}{|x-y|}\cdot\delta x+\frac{|y|}{|x-y|}\cdot\delta y.$$

Пусть x и y — слагаемые одного знака, а $\delta = max(\delta x, \delta y)$. Тогда

$$\delta(x+y)\leq\frac{\delta(|x|+|y|)}{|x+y|}=\delta,$$

т.е. $\delta(x+y)=\delta$

Если приближенные слагаемые имеют один знак, то граница относительной погрешности их суммы не превышает наибольшей из границ относительных погрешностей слагаемых.

Катастрофическая потеря точности

При вычитании близких чисел может произойти большая потеря точности. В случае, если вычитаемые числа почти одинаковы, то даже при условии, что их собственные ошибки малы, относительная погрешность разности может оказаться большой.
Пример. Найдем разность чисел $x = 62,425$ и $y = 62,409$, у которых все цифры верны с строгом смысле.

$$x - y = 62,425 - 62,409 = 0,016.$$

Граница абсолютной погрешности разности

$$\Delta(x - y) = 0,0005 + 0,0005 = 0,001,$$

поэтому в числе $0,016$ из двух значащих цифр верна лишь одна.

Сравним границы относительных погрешностей результата и исходных данных:

$$\delta x = \frac{0,0005}{62,425} = 0,000008$$

$$\delta x = \frac{0,0005}{62,409} = 0,000008$$

$$\delta(x-y) = \frac{0,001}{0,016} = 0,0625.$$

Таким образом, граница относительной погрешности разности оказалась пости в 8 тысяч раз больше границы относительной погрешности исходных данных. Поэтому в приближенных вычислениях лучше исключать вычитание близких по величине значений (например, путем преобразования вычисляемых выражений).

Пример. Ошибка вызванная точностью представления числа в формате IEEE 754

Листинг: Код на C++

```
1  //Потеря точности
2  #include <stdio.h>
3  using namespace std;
4
5  int main() {
6      float a, b, f;
7      a=123456789;
8      b=123456788;
9      f=a-b;
10     printf("Result: %f\n", f);
11     return 0;
12 }
13
14 Result: 8.000000
```

Ответ должен быть 1.000000.

Если пример решить аналитически, то ответ будет 1.
Абсолютная ошибка равна +7.
Почему ответ получился неправильным?
Число 123456789 в single=4CEB79A3hex(ieee)=123456792(dec) абсолютная ошибка представления равна +3
Число 123456788 в single=4CEB79A2hex(ieee)=123456784(dec) абсолютная ошибка представления равна -4
Относительная погрешность исходных чисел приблизительно равна 3,24e-6%
В результате одной операции относительная погрешность результата стала 800%, т.е. увеличилась в 2,5e+8 раз.
Здесь произошло катастрофическое понижение точности вычислений в операции где абсолютное значение результата много меньше любого из входного значений переменных.

Произведение и частное

Пусть $p = x \cdot y$ — произведение двух приближенных чисел, а $q = \frac{x}{y}$ — их частное. Знаки чисел x и y не влияют на величину ошибки, поэтому для простоты примем, $x, y > 0$.

$$lnp = lnx + lny, lnq = lnx - lny.$$

Известно, что

$$df(x) \approx \Delta f(x),$$

Произведение и частное

Тогда

$$\Delta lnz \approx dlnz = \frac{\Delta z}{z}.$$

$$\Delta lnp = \Delta lnq = \Delta lnx + \Delta lny,$$

$$\frac{\Delta p}{p} = \frac{\Delta q}{q} = \frac{\Delta x}{x} + \frac{\Delta y}{y}, \tag{3}$$

Следовательно,

$$\delta(x \cdot y) = \delta(\frac{x}{y}) = \delta x + \delta y.$$

Из формул (3) легко получаются формулы

$$\Delta(xy) = x \cdot \Delta y + y \cdot \Delta x,$$

$$\Delta\left(\frac{x}{y}\right) = \frac{x \cdot \Delta y + y \cdot \Delta x}{y^2},$$

Вычисление погрешностей значений элементарных функций

Пусть функция $f(x)$ дифференцируема в некоторой окрестности x, а e_x — абсолютная ошибка значения аргумента. Тогда

$$e_f = |f(x + \Delta x) - f(x)|.$$

Так как на практике ошибка e_x обычно мала по сравнению со значением x, воспользуемся приближенным равенством $e_f \approx |df| = |f'(x)| \cdot e_x$. Заменим e_x на Δx :

$$e_f \leq |f'(x)| \cdot \Delta x.$$

Тогда

$$\Delta f = |f'(x)| \cdot \Delta x.$$

Вычисление погрешностей значений элементарных функций

Пример. Определить количество верных в строгом смысле цифр в значении функции $f(x) = \sin^2 x$ при $x = 2,321$. В числе x все цифры верны в строгом смысле.

Решение.

$$f'(x) = \sin 2x, \ \Delta x = 0,0005.$$

$$\Delta f(x) = |\sin(2 \cdot 2,321)| \cdot 0,0005 \approx 0.0004987618591 \leq$$

$$\leq 0.0005.$$

Таким образом, в значении $sin^2(2,321) \approx 0.535165434904$ верны в строгом смысле 4 цифры.