МЕХАНИКО-МАТЕМАТИЧЕСКИЙ И ХИМИЧЕСКИЙ ФАКУЛЬТЕТЫ МГУ ИМ. М.В.ЛОМОНОСОВА

МЕЖДИСЦИПЛИНАРНЫЕ НАУЧНО-ОБРАЗОВАТЕЛЬНЫЕ ШКОЛЫ МГУ

ПРИБЛИЖЁННЫЕ ВЫЧИСЛЕНИЯ

В.Г.ЧИРСКИЙ

2022

Рекомендовано методической комиссией химического факультета и кафедрой математического анализа механико-математического факультета МГУ им. М.В. Ломоносова в качестве учебного пособия для студентов

В современном мире компьютерные технологии внедрены практически во все научные и прикладные исследования. В свою очередь, это вызывает бурное развитие математических дисциплин, поскольку многие из математических дисциплин имеют важное прикладное значение и являются основой компьютерных технологий.

Для правильного и эффективного использования многих математических программ требуется умение сформулировать задачи, возникающие в процессе исследования математической модели изучаемого явления, выбрать подходящий алгоритм решения, осмыслить полученный результат. Для этого требуется достаточный уровень математической подготовки.

В серии методических разработок «математика для современной химии» в рамках проекта «Междисциплинарные научно-образовательные школы МГУ» рассматриваются вопросы, усвоение которых способствует повышению математической культуры учащихся, развитию их профессиональных компетенций. Выбор тем разработок не случаен. Он основан на методических исследованиях кафедры математического анализа, на учёте мнений кафедр химического факультета, на анализе результатов экзаменов.

Важная цель этих разработок – облегчить самостоятельную работу студентов и способствовать успешной сдаче экзаменов и зачётов. В этом пособии содержится материал, относящийся к курсу математического анализа, читаемому студентам первого курса химического факультета МГУ.

Приближённые вычисления

Как известно, действительные числа изображаются конечными или бесконечными десятичными дробями. Поскольку принципиально невозможно производить практические операции над бесконечными дробями, их заменяют приближёнными значениями этих чисел.

Обычно приближённым значением a действительного числа A называется число, незначительно отличающееся от числа A и заменяющее это число A в вычислениях. Чтобы сделать результаты приближённых вычислений надёжными, следует соблюдать правила приближённых вычислений.

Эта методическая разработка посвящена правилам приближённых вычислений, часто используемым на практике.

Определение. Под ***ошибкой*** или ***погрешностью*** Δa приближённого числа a обычно понимают разность между соответствующим точным числом A и числом a, т.е. $\Delta a = A - a$. Удобно рассматривать ***абсолютную погрешность*** Δ приближённого числа a, равную абсолютной величине погрешности Δa, т.е. $\Delta = |\Delta a|$.

Часто бывает так, что точное значение числа A неизвестно. В этом случае абсолютная погрешность также неизвестна, и следует попытаться найти оценку сверху для этой погрешности, т.е. такое число Δ_a, про которое известно, что оно не меньше, чем Δ. Тогда неравенство $\Delta \leq \Delta_a$ означает, что $|A - a| \leq \Delta_a$, или, что то же самое, что $a - \Delta_a \leq A \leq a + \Delta_a$. Таким образом, число $a - \Delta_a$ будет являться приближением для числа A по недостатку, а число $a + \Delta_a$ будет являться приближением для числа A по избытку. Неравенства $a - \Delta_a \leq A \leq a + \Delta_a$. часто кратко записывают так: $A = a \pm \Delta_a$.

Разумеется, если $\tilde{\Delta} \geq \Delta_a$, то из приближённого равенства $A = a \pm \Delta_a$ следует менее точное приближённое равенство $A = a \pm \tilde{\Delta}$.

Число Δ_a, т.е. точность приближения, выбирается, в основном, исходя из потребностей решаемой задачи. Это означает, что одно и то же число может быть приближено с различной точностью (примеры будут даны ниже). Таким образом, вместо точного значения действительного числа A мы рассматриваем совокупность его различных приближённых значений с различными заданными точностями.

Определение ***Относительная погрешность*** δ приближённого числа a, по определению, равна $\delta = \frac{|A-a|}{|A|}$.

В тех случаях, когда точное значение числа A неизвестно, неизвестно и точное значение числа δ. Тогда следует получить оценку сверху для относительной погрешности, исходя из оценки сверху для абсолютной погрешности. Например, при условиях $A > 0, a > \Delta_a$ справедлива оценка:

$$\delta \leq \delta_a = \frac{\Delta_a}{a-\Delta_a}.$$

Из курса средней школы известно, что любое рациональное число можно представить в виде конечной или периодической бесконечной десятичной дроби. Остальные действительные числа (т.е. иррациональные числа) изображаются бесконечными непериодическими десятичными дробями.

На практике действия с бесконечными дробями невозможны и их приходится заменять действиями с конечными десятичными дробями, служащими приближениями для рассматриваемых чисел, т.е. с числами вида a или $-a$, где

$$a = a_m 10^m + \cdots + a_{m-n+1} 10^{m-n+1}, a_m \neq 0$$

и $a_i, i = 1, \dots, m-n+1 -$ десятичные цифры.

Определение В приведённом представлении числа a ***значащими цифрами*** называются все отличные от нуля цифры, все те равные нулю цифры, которые содержатся между отличными от нуля значащими цифрами, а также равные нулю цифры, необходимые для обозначения десятичных разрядов целого числа. Говорят, что n значащих цифр приближённого числа являются

верными, если абсолютная погрешность этого приближённого числа a не превышает половины единицы разряда, выражаемого n -ой значащей цифрой, считая слева направо.

Таким образом, если для приближённого числа a, заменяющего точное число A, выполняется неравенство $|A-a| \leq \frac{1}{2}10^{m-n+1}$, то, по определению, первые n значащих цифр этого числа являются верными.

Здесь будут уместно следующее замечание. Во многих случаях верные знаки приближающего числа совпадают с соответствующими цифрами точного числа, например, для точного числа $A = 411{,}23$ приближённое число $a = 411{,}20$ имеет четыре верных знака, так как $|A-a| = 0{,}03 < \frac{1}{2}0{,}1$, причём эти знаки совпадают со знаками точного числа, но для точного числа $A = 37{,}28$ приближённое число $a = 37{,}30$ имеет три верных знака, так как $|A-a| = 0{,}02 < \frac{1}{2}0{,}1$, а совпадают только две цифры. В примере $A = 100, a = 99{,}9, |A-a| = 0{,}1 < \frac{1}{2}$ у приближённого числа имеется два верных знака, ни один из которых не совпадает со знаками исходного числа.

Округление точного или приближённого числа состоит в замене его числом, имеющим меньшее количество значащих цифр. Действительно, точность приближения определяется не всеми значащими цифрами, а только верными.

Обычно используют такое практическое правило: при выполнении приближённых вычислений число значащих цифр промежуточных результатов число не должно превосходить числа верных цифр более чем на две единицы.

Чтобы округлить число до n значащих цифр все его цифры, расположенные правее n -ой значащей цифры либо отбрасывают, либо заменяют нулями в случае, когда это необходимо для сохранения разрядов целого числа. Кроме того, если первая из отброшенных цифр меньше 5, то предыдущие разряды не меняются. Если первая из отброшенных цифр больше

пяти, то к последней оставшейся цифре прибавляют единицу. Если первая из отброшенных цифр равна пяти, причём среди остальных отброшенных цифр есть не равные нулю, то к последней оставшейся цифре прибавляют единицу. Если первая из отброшенных цифр равна пяти, причём все остальные отброшенные цифры равны нулю, то к последней оставшейся цифре добавляют единицу, если она нечётная и не меняют её, если она чётная. Приведём примеры округления до трёх значащих цифр:

$$123456789 \cong 123000000, 34{,}567 \cong 34{,}6, 12{,}2500 \cong 12{,}2, 12{,}350 \cong 12{,}4.$$

(*Здесь символ* $\cong$ *обозначает знак приближённого равенства*).

Можно доказать следующее утверждение:

Теорема. ***Если положительное приближённое число***

$$a = a_m 10^m + \cdots + a_{m-n+1} 10^{m-n+1}, a_m \neq 0,$$

имеет n ***верных десятичных знаков, то относительная погрешность*** δ ***этого числа не превосходит величины***

$$\delta \leq \frac{1}{a_m}(0{,}1)^{n-1}.$$

Доказательство. По определению,

$$\delta_a = |A - a| \leq \frac{1}{2} 10^{m-n+1}.$$

Значит,

$$A \geq a - \frac{1}{2} 10^{m-n+1} \geq a_m 10^m - \frac{1}{2} 10^{m-n+1} = \frac{1}{2} 10^m \left(2a_m - \frac{1}{10^{n-1}}\right) \geq$$

$$\frac{1}{2} 10^m (2a_m - 1) \geq \frac{1}{2} a_m 10^m$$

и

$$\delta = \frac{|A-a|}{|A|} \leq \frac{\frac{1}{2}10^{m-n+1}}{\frac{1}{2}a_m 10^m} = \frac{1}{a_m}(0{,}1)^{n-1},$$

что и требовалось доказать.

Теорема. ***Абсолютная погрешность алгебраической суммы нескольких приближённых чисел не превышает суммы абсолютных погрешностей слагаемых.***

Доказательство. Пусть $A_1, \dots, A_k -$ точные значения, $a_1, \dots, a_k -$ приближающие их числа. Тогда

$$\left|\left((A_1 + \cdots + A_k) + \cdots + A_k\right) - (a_1 + \cdots + a_k)\right| \leq |A_1 - a_1| + \cdots + |A_k - a_k|$$

по известному свойству абсолютной величины, что и требовалось доказать.

Это утверждение означает, что

$$\Delta_{a_1+\cdots+a_k} \leq \Delta_{a_1} + \cdots + \Delta_{a_k}.$$

Поэтому обычно правую часть этого неравенства и принимают за оценку абсолютной погрешности суммы. Таким образом, абсолютная погрешность суммы оказывается не меньше, чем наибольшая из абсолютных погрешностей слагаемых. Следовательно, не имеет смысла сохранять излишние знаки и в более точных слагаемых.

Итак, при сложении приближённых чисел используется такое простое правило. Во-первых, следует найти числа, десятичная запись которых содержит наименьшее количество знаков после запятой. Остальные числа округлить так же, как найденные выше, взяв ещё один лишний знак. Произвести сложение полученных округлённых чисел и округлить результат.

Теорема. ***Относительная погрешность суммы слагаемых одного и того же знака не превышает наибольшей относительной погрешности.***

Доказательство. Пусть, как и выше, точные числа равны $A_1, \dots, A_k$, приближённые числа равны $a_1, \dots, a_k$, абсолютные погрешности оценены числами $\Delta_1, \dots, \Delta_k$, относительные погрешности равны $\delta_1, \dots, \delta_k$ и наибольшая из них есть $\tilde{\delta}$. Тогда относительная погрешность суммы не превосходит числа

$$\delta \leq \frac{\Delta_1+\cdots+\Delta_k}{|A_1|+\cdots+|A_k|} \leq \frac{\delta_1|A_1|+\cdots+\delta_k|A_k|}{|A_1|+\cdots+|A_k|} \leq \tilde{\delta}.$$

Сформулируем следующую теорему без доказательства.

Теорема. ***Относительная погрешность произведения нескольких положительных приближённых чисел не превышает суммы их относительных погрешностей.***

Полезно руководствоваться таким правилом. Пусть мы ищем произведение нескольких приближённых сомножителей. Тогда, во-первых, округлим все сомножители, кроме наименее точного, так, чтобы они имели на одну значащую цифру больше, чем число верных цифр в этом наименее точном из сомножителей. В результате умножения сохранить столько значащих цифр, сколько верных цифр в наименее точном из сомножителей.

Как определить число верных знаков произведения? Рассмотрим $k \leq 10$ сомножителей $a_1, \dots, a_k$, каждый из которых пусть имеет $n > 1$ верных цифр. Пусть $\tilde{a}_1, \dots, \tilde{a}_k$ - их первые значащие цифры. Тогда, как доказано выше,

$\delta_i \leq \frac{1}{2\tilde{a}_i}(0{,}1)^{n-1}$,

и по предыдущему утверждению,

$$\delta \leq \frac{1}{2}\left(\frac{1}{\tilde{a}_1} + \dots + \frac{1}{\tilde{a}_k}\right)(0{,}1)^{n-1}.$$

Поскольку, по условию, $k \leq 10$ и все $\tilde{a}_i \geq 1$, получаем, что

$$\delta \leq \frac{1}{2}(0{,}1)^{n-2}.$$

Следовательно, число верных знаков может уменьшиться не более, чем на 2. Если сомножители имеют разное количество верных цифр, то под числом $n, n > 1$ следует понимать наименьшее из чисел верных знаков сомножителей.

Вопрос о погрешности частного решается примерно так же, как и в случае произведения. Именно,

относительная погрешность частного не
превосходит суммы относительных погрешностей делимого и делителя.

Наконец, обратимся к операциям возведения в натуральную степень m и извлечения корня степени m. В первом случае:

оценка относительной погрешности m -ой степени числа в m раз больше, чем оценка относительной погрешности самого числа.

Для корня имеем:

оценка относительной погрешности корня m -ой степени из положительного числа в m раз меньше, чем оценка относительной погрешности самого числа.

Изложенное выше даёт лишь начальное представление о методах приближённых вычислений. Для исследования задачи приближённого вычисления значений функций общего вида сначала придётся изучить основные понятия дифференциального исчисления и доказать формулы Тейлора. Об этом будет подробнее сказано в следующих разработках. Сейчас же мы рассмотрим примеры применения указанных выше правил.

Пример 1.Сложим числа $2{,}173 \pm 0{,}0005, 0{,}11 \pm 0{,}005, 43{,}1244 \pm 0{,}00005$. Ясно, что точность вычисления определяется вторым слагаемым. Поэтому, в соответствии с выписанным выше правилом, сохраним первое и второе числа и округлим третье следующим образом: $43{,}124 \pm 0{,}0005$. Тогда первое и третье слагаемые дадут в сумме $45{,}297 \pm 0{,}001$ $45,297 \pm 0,001$. Добавление второго слагаемого приведёт к $45{,}41 \pm 0{,}006$. Из этого следует, что верными цифрами суммы будут первые три её цифры.

Пример 2. Найдём произведение приближённых чисел

$11{,}3 \pm 0{,}05, 28{,}46 \pm 0{,}005.$

Оценкой сверху для относительной погрешности служит число.$\delta = \frac{0{,}05}{11{,}25} + \frac{0{,}005}{28{,}455} \leq 0{,}0045.$

Произведение приблизительно равно 323,08 и абсолютная погрешность не превосходит 1,45. Поэтому произведение имеет два верных знака и его следует записать так: 323 ± 2.

Пример 3. Сколько десятичных знаков числа $\sqrt{22}$ следует взять, чтобы относительная погрешность вычисления не превышала $0{,}001$?.

Решение: Первая цифра этого числа, очевидно, равна 4. Для того, чтобы выполнялась оценка, $\delta = \frac{1}{4}10^{-n+1} \leq 0{,}001$достаточно взять $n = 4$.

Задача. Оценить относительную погрешность замены числа π числом 3,14.

Решение. $a_m = 3, n = 3$, поэтому $\delta \leq \frac{1}{3}10^{-2}$. (На самом деле в этом примере можно дать и более точную оценку, так как мы знаем, что $|\pi - 3{,}14| < 0{,}0016, \frac{0{,}0016}{3{,}14} < 0{,}00051$).

Рассмотрим такой интересный пример. Первым приближением, известным ещё Архимеду в III веке до н.э. для числа π, равного отношению длины окружности к её диаметру, служит число $\frac{22}{7} = 3{,}142857\ldots$. Зная разложение числа $\pi = 3{,}1415926\ldots$, получаем, что $\left|\pi - \frac{22}{7}\right| < 0{,}0013 < 0{,}5\ 10^{-2}$, что означает, по определению, что три значащих цифры этого приближённого значения числа π являются верными.

Адриан Меций, голландский геометр XVI века, предложил для приближения числа π число $\frac{355}{113}$; это число легко запомнить по правилу: написав по два раза нечётные цифры 1,1,3,3,5,5, следует последние три взять цифрами числителя, а первые три - знаменателя. Так как $\frac{355}{113} = 3{,}1415929\ldots$, а как отмечалось выше, $\pi = 3{,}1415926\ldots$, то $\left|\pi - \frac{355}{113}\right| < 5 \cdot 10^{-7}$и приближённое значение $\frac{355}{113}$ для числа π имеет 7 верных знаков.

Важным направлением развития современной вычислительной математики являются разработка и реализация алгоритмов, дающих огромные количества верных знаков в десятичном разложении числа π. В 2002 году их было известно уже $1{,}24 \cdot 10^{12}$! Разумеется, такие вычисления не являются самоцелью или демонстрацией вычислительных возможностей современных компьютеров. Знание такого большого количества десятичных знаков числа π предоставляет огромный выбор псевдослучайных чисел, широко применяемых при вероятностных расчётах и др. (Интересно, что среди первого миллиона цифр все десятичные цифры встречаются примерно с одинаковой частотой).