Лекция 2

Тема 2. Основные понятия элементарной теории погрешностей

§ 2.1. Источники и классификация погрешностей результатов численного решения задач. Как отмечалось в первой лекции, при решении прикладной задачи с использованием ЭВМ получить точное решение задачи практически невозможно. Получаемое решение почти всегда содержит погрешность, т.е. является приближенным.

Наличие погрешности решения обусловлено рядом причин. Перечислим их.

1). При решении задач на ЭВМ имеют дело, как правило, с математической моделью физического явления, которая является его приближенным описанием. В связи с этим получаемые в результате вычислений характеристики процесса или явления содержат погрешность, величина которой зависит от степени адекватности модели реальному процессу.

2). Исходные данные, как правило, содержат погрешности, поскольку они получаются в результате измерений либо являются результатом решения некоторых задач.

3). Применяемые при решении задачи методы, как отмечено в первой лекции, в большинстве случаев являются приближенными.

4). При вводе исходных данных в ЭВМ, выполнении арифметических операций и выводе результатов на печать или дисплей монитора производятся округления, определяемые разрядностями устройств ввода/вывода и ЭВМ.

Пусть y – точное значение величины, вычисление которой является целью поставленной задачи, а y^* – ее приближенное значение. Соответствующая первым двум из указанных причин погрешность $\delta_{\text{н}} y^*$ называется *неустранимой погрешностью.* Такое название обусловлено тем, что математическая модель и исходные данные вносят в решение ошибку, которая не может быть устранена далее. Единственный способ уменьшить эту погрешность – перейти к более точной математической модели и задать более точные исходные данные. Но это не всегда возможно.

Погрешность $\delta_{\text{м}}y^*$, источником которой является метод решения задачи, называется *погрешностью метода,* а погрешность $\delta_{\text{в}}y^*$, возникающая при вводе, выводе и вычислениях, – *вычислительной погрешностью.*

Таким образом, полная погрешность результата решения задачи на ЭВМ $\delta y^* = y - y^*$ складывается из трех составляющих: неустранимой погрешности, погрешности метода и вычислительной погрешности, т.е. $\delta y^* = \delta_{\text{н}}y^* + \delta_{\text{м}}y^* + \delta_{\text{в}}y^*$.

Будем считать, что математическая модель фиксирована и входные данные задаются извне, так что повлиять на значение величины $\delta_{\text{н}}y^*$ в процессе решения задачи нельзя. Однако это не означает, что предварительные оценки величины неустранимой погрешности не нужны. Достоверная информация о порядке величины $\delta_{\text{н}}y^*$ позволяет осознанно выбрать метод решения задачи и разумно задать его точность. На практике исходят из того, что погрешность метода должна быть на порядок (в 2 – 10 раз) меньше неустранимой погрешности. Большее значение $\delta_{\text{м}}y^*$ощутимо снижает точность результата, меньшее – требует увеличения затрат, практически мало влияя на значение полной погрешности.

Величина вычислительной погрешности в основном определяется характеристиками используемой ЭВМ. Желательно, чтобы величина $\delta_{\text{в}}y^*$ была хотя бы на порядок меньше величины погрешности метода.

§ 2.2. Приближенные числа. Абсолютная и относительная погрешности. Итак, как отмечено выше, числа, получаемые при решении на ЭВМ прикладных задач, являются приближенными. Следовательно, вопрос о точности результатов, т.е. о мере их уклонения от истинных значений, в теории и практике приближенных вычислений приобретает особое значение. Начнем его рассмотрение с введения основных понятий элементарной теории погрешностей.

1. Абсолютная и относительная погрешности. Пусть a – точное (неизвестное) значение некоторой величины, a^* – приближенное (известное) значение той же величины (*приближенное число*). *Ошибкой (погрешностью)* приближенного числа a^* называется разность $a - a^*$ между точным и приближенным значениями.

Простейшей количественной мерой ошибки является *абсолютная погрешность*

$$\Delta(a^*) = |\, a - a^*|. \qquad (2.1)$$

Однако по величине абсолютной погрешности не всегда можно сделать правильное заключение о качестве приближения. Действительно, если $\Delta(a^*) = 0.1$, то следует ли считать точность большой или малой? Ответ существенным образом зависит от принятых единиц измерения и масштабов величин. Например, если $a \approx 0.3$, то точность приближения невысока. Если же $a \approx 3\cdot10^8$, то точность следует признать очень высокой. Таким образом, естественно соотнести погрешность величины и ее значение, для чего вводят понятие *относительной погрешности* (при *a ≠ 0)*

$$\delta(a^*) = \frac{|\, a - a^*|}{|a|} = \frac{\Delta(a^*)}{|a|}\ . \qquad (2.2)$$

Очевидно, что относительные погрешности не зависят от масштабов величин. В частности, для приведенного выше примера $\delta(a^*) \approx 0.33 = 33\%$ в первом случае и $\delta(a^*) \approx 0.33\cdot10^{-9} = 0.33\cdot10^{-7}\%$ во втором.

Так как значение *a* неизвестно, то непосредственное вычисление величин $\Delta(a^*)$ и $\delta(a^*)$ по формулам (1) и (2) невозможно. Более реальная и часто поддающаяся решению задача состоит в получении оценок вида

$$|\, a - a^*| \le \overline{\Delta}(a^*), \qquad (2.3)$$

$$\frac{|a - a^*|}{|a|} \le \overline{\delta}(a^*), \qquad (2.4)$$

где $\overline{\Delta}(a^*)$ и $\overline{\delta}(a^*)$ – известные величины, которые мы будем называть *верхними границами* (или просто *границами*) *абсолютной* и *относительной погрешностей*.

Если величина $\overline{\Delta}(a^*)$ известна, то неравенство (4) будет выполнено, если положить

$$\overline{\delta}(a^*) = \frac{\overline{\Delta}(a^*)}{|a|}. \qquad (2.5)$$

Точно так же если величина $\bar{\delta}(a^*)$ известна, то следует положить

$$\bar{\Delta}(a^*) = |a|\,\bar{\delta}(a^*) \qquad (2.6)$$

Поскольку значение a неизвестно, при практическом применении формулы (5), (6) заменяют приближенными равенствами

$$\bar{\delta}(a^*) \approx \frac{\bar{\Delta}(a^*)}{|a^*|},\ \bar{\Delta}(a^*) \approx |a^*|\bar{\delta}(a^*) \qquad (2.7)$$

2. Правила записи приближенных чисел. Пусть приближенное число a^* задано в виде конечной десятичной дроби:

$$a^* = \alpha_n\alpha_{n-1}\ldots\alpha_0.\,\beta_1\beta_2\ldots\beta_m.$$

Значащими цифрами числа a^* называют все цифры в его записи, начиная с первой ненулевой слева.

Пример 1. У чисел a^*= 0.0<u>103</u> и a^* = 0.0<u>103000</u> значащие цифры подчеркнуты.

Значащую цифру называют *верной*, если абсолютная погрешность числа не превосходит единицы разряда, соответствующего этой цифре.

Пример 2. Если $\bar{\Delta}(a^*) = 2\cdot10^{-6}$, то число a^* = 0.0<u>1030</u>00 имеет 4 верные значащие цифры (они подчеркнуты).

Распространенной ошибкой является отбрасывание последних значащих нулей (даже если они представляют верные цифры).

З а м е ч а н и е. Верная цифра приближенного числа, вообще говоря, не обязана совпадать с соответствующей цифрой в записи точного числа. Таким образом, термин «верная цифра» не следует понимать буквально.

Пример 3. Пусть a = 1.00000, a^*= 0.99999. Тогда $\Delta(a^*)$=0.00001 и у числа a^*= 0.<u>99999</u> все подчеркнутые цифры верные, хотя они и не совпадают с соответствующими цифрами числа a.

Если число a^* имеет ровно N верных значащих цифр, то $\delta(a^*) \sim 10^{-N}$.

Заметим, что границы абсолютной и относительной погрешностей принято записывать с одной или двумя значащими цифрами. Бóльшая точность в записи этих величин не имеет смысла, так как обычно они

являются довольно грубыми оценками истинных значений погрешностей.

Пример 6. Информация о погрешности вида $\delta(a^*) \approx 0.288754 \cdot 10^{-5}$ практически равноценна информации $\delta(a^*) \approx 3 \cdot 10^{-6}$, причем последняя вызывает больше доверия. Скорее всего, вполне удовлетворительной в данном случае является запись $\delta(a^*) \sim 10^{-6}$.

Вернемся к неравенству (3). Очевидно, что оно эквивалентно двойному неравенству

$a^* - \bar{\Delta}(a^*) \leq a \leq a^* + \bar{\Delta}(a^*)$

и поэтому тот факт, что число a^* является приближенным значением числа a с верхней границей абсолютной погрешности $\bar{\Delta}(a^*)$ (или с *абсолютной точностью* $\varepsilon = \bar{\Delta}(a^*)$) принято записывать в виде $a = a^* \pm \bar{\Delta}(a^*)$. Как правило, числа a^* и $\bar{\Delta}(a^*)$ указывают с одинаковым числом цифр после десятичной точки.

Пример 7. Пусть для числа a известны приближенное значение $a^* = 1.648$ и граница абсолютной погрешности $\bar{\Delta}(a^*) = 0.002832$. Тогда можно записать $a = 1.648 \pm 0.003$.

Рассмотрим теперь неравенство (4). Из него следует, что значение a заключено примерно между $a^*(1 - \bar{\delta}(a^*)$ и $a^*(1 + \bar{\delta}(a^*)$. Поэтому тот факт, что число a^* является приближенным значением числа a с границей относительной погрешности $\bar{\delta}(a^*)$ (или с *относительной точностью* $\varepsilon = \bar{\delta}(a^*)$) принято записывать в виде $a = a^*(1 \pm \bar{\delta}(a^*)$.

Пример 8. Оценим точность приближения π^*=3.14 к числу π. Известно, что $\pi = 3.14159\ldots$, поэтому $\pi - \pi^* = 0.00159\ldots$ Следовательно, можно принять $\Delta(\pi^*) = 0.0016$ и $\bar{\delta}(\pi^*) \approx 0.0016 / 3.14 \approx 0.00051 = 0.051\%$. Итак, $\pi = 3.14\ (1 \pm 0.051\%)$.

З а м е ч а н и е. Если число a^* приводится в качестве результата без указания величины погрешности, то принято считать, что все его значащие цифры являются верными. Начинающий пользователь ЭВМ часто слишком доверяет выводимым из ЭВМ цифрам, предполагая, что вычислительная машина придерживается того же соглашения. Однако это совсем не так: число может быть выведено с таким количеством

значащих цифр, сколько потребует программист заданием соответствующего формата. Как правило, среди этих цифр только небольшое число первых окажутся верными, а возможно, что верных цифр нет совсем. Анализировать результаты вычислений и определять степень их достоверности совсем непросто. Одна из целей изучения вычислительных методов и состоит в достижении понимания того, что можно и чего нельзя ожидать от результатов, полученных на ЭВМ.

3. Округление. Часто возникает необходимость в *округлении* числа a, т.е. в замене его другим числом a^* с меньшим числом значащих цифр. Возникающая при такой замене погрешность называется *погрешностью округления.*

Существует несколько способов округления числа до n значащих цифр. Наиболее простой из них – *усечение* состоит в отбрасывании всех цифр, расположенных справа от n-й значащей цифры. Более предпочтительным является *округление по дополнению*. В простейшем случае это правило состоит в следующем. Если первая слева от отбрасываемых цифр меньше 5, то сохраняемые цифры остаются без изменения. Если же она больше либо равна 5, то в младший сохраняемый разряд добавляется единица.

Абсолютная величина погрешности при округлении по дополнению не превышает половины единицы разряда, соответствующего последней оставляемой цифре, а при округлении усечением – единицы того же разряда.

Пример 9. При округлении числа a = 1.72631 усечением до трех значащих цифр получится число a^*=1.72, а при округлении по дополнению число a^*=1.73.

Границы абсолютной и относительной погрешностей принято округлять в сторону увеличения.

Пример 10. Округление по дополнению до двух значащих цифр величин $\bar{\Delta}(a^*)$ = 0.003721 и $\bar{\delta}(a^*)$ = 0.0005427 дает значения $\bar{\Delta}(a^*)$ = 0.0038 и $\bar{\delta}(a^*)$ = 0.00055.

§ 2.3. Погрешности арифметических операций над приближенными числами

Исследуем влияние погрешностей исходных данных на погрешность результатов арифметических операций. Пусть a^* и b^* - приближенные значения чисел a и b. Какова соответствующая им величина неустранимой погрешности результата?

Т е о р е м а 1. *Абсолютная погрешность алгебраической суммы или разности не превосходит суммы абсолютных погрешностей слагаемых, т.е.*

$$\Delta(a^* \pm b^*) \le \Delta(a^*) + \Delta(b^*). \qquad (2.8)$$

□ *Имеем* $\Delta(a^* \pm b^*) = |(a \pm b) - (a^* \pm b^*)| = |(a - a^*) \pm (b - b^*)| \le \Delta(a^*) + \Delta(b^*)$ ■

С л е д с т в и е. В силу неравенства (8) естественно положить

$$\bar{\Delta}(a^* \pm b^*) = \bar{\Delta}(a^*) + \bar{\Delta}(b^*). \qquad (2.9)$$

Оценим относительную погрешность алгебраической суммы и разности.

Т е о р е м а 2. *Пусть a и b ненулевые числа одного знака. Тогда справедливы неравенства*

$$\delta(a^* + b^*) \le \delta_{max}, \quad \delta(a^* - b^*) \le \nu\,\delta_{max}, \qquad (2.10)$$

где $\delta_{max} = max\{\delta(a^*), \delta(b^*)\}$, $\nu = \frac{|a+b|}{|a-b|}$.

□ Используя формулу (2.2) и неравенство (2.8) имеем

$|a \pm b|\,\delta(a^* \pm b^*) = \bar{\Delta}(a^* \pm b^*) \le \Delta(a^*) + \Delta(b^*) = |a|\,\delta(a^*) + |b|\,\delta(b^*) \le (|a| + |b|)\,\delta_{max} = |a + b|\,\delta_{max}$.

Из полученного неравенства следуют оценки (2.10). ■

С л е д с т в и е. *В силу неравенств (2.10) естественно положить*

$$\bar{\delta}(a^* + b^*) = \bar{\delta}_{max}, \quad \bar{\delta}(a^* - b^*) = \nu\,\bar{\delta}_{max}, \qquad (2.11)$$

где $\bar{\delta}_{max} = max\{\bar{\delta}(a^*), \bar{\delta}(b^*)\}$, $\nu = \frac{|a+b|}{|a-b|}$.

Первое из неравенств (2.11) означает, что при суммировании чисел одного знака не происходит потери точности, если точность оценивать в относительных единицах. Совсем иначе обстоит дело при вычитании чисел одного знака. Здесь граница относительной погрешности возрастает в $\nu > 1$ раз и возможна существенная потеря точности. Если числа a и b близки настолько, что $|a + b| >> |a - b|$, то $\nu >> 1$

и не исключена полная или почти полная потеря точности. Когда это происходит, говорят о том, что произошла *катастрофическая потеря точности.*

Итак, получаем следующий важный вывод. При построении численного метода решения задачи следует избегать вычитания близких чисел одного знака. Если же такое вычитание неизбежно, то следует вычислять аргументы с повышенной точностью, учитывая ее потерю примерно в $\nu = \frac{|a+b|}{|a-b|}$ раз.

Т е о р е м а 3. *Для относительных погрешностей произведения и частного приближенных чисел верны оценки*

$$\delta(a^*b^*) \le \delta(a^*) + \delta(b^*) + \delta(a^*)\,\delta(b^*), \qquad (2.12)$$

$$\delta\left(\frac{a^*}{b^*}\right) \le \frac{\delta(a^*) + \delta(b^*)}{1-\delta(b^*)}, \qquad (2.13)$$

в последней из которых считается, что $\delta(b^*) < 1$.

□ Выполним следующие преобразования:

$|ab|\,\delta(a^*b^*) = \Delta(a^*b^*) = |ab - a^*b^*| = |(a-a^*)b + (b-b^*)a - (a-a^*)(b-b^*)| \le |b|\,\Delta(a^*) + |a|\,\Delta(b^*) + \Delta(a^*)\,\Delta(a^*) = |ab|(\delta(a^*) + \delta(b^*) + \delta(a^*)\delta(b^*))$,

т.е. $|ab|\,\delta(a^*b^*) \le |ab|(\delta(a^*) + \delta(b^*) + \delta(a^*)\delta(b^*))$.

Разделив обе части этого неравенства на $|ab|$, получаем оценку (2.12).

Для вывода второй оценки предварительно заметим, что $|b^*| = |b + (b^* - b)| \ge |b| - \Delta(b^*) = |b|\,(1 - \delta(b^*))$. Тогда

$$\delta\left(\frac{a^*}{b^*}\right) = \frac{\left|\frac{a}{b} - \frac{a^*}{b^*}\right|}{\left|\frac{a}{b}\right|} = \frac{|ab^* - b\,a^*|}{|ab^*|} = \frac{|a(b^*-b) + b\,(a-a^*)|}{|ab^*|} \le \frac{|a|\Delta(b^*) + |b|\Delta(a^*)}{|ab|(1-\delta(b^*))} = \frac{\delta(a^*) + \delta(b^*)}{1-\delta(b^*)}.$$ ■

С л е д с т в и е. Если $\delta(a^*) << 1$ и $\delta(b^*) << 1$, то для оценки границ относительных погрешностей можно использовать следующие приближенные равенства:

$$\bar{\delta}(a^*b^*) \approx \bar{\delta}(a^*) + \bar{\delta}(b^*), \quad \bar{\delta}(a^*/b^*) \approx \bar{\delta}(a^*) + \bar{\delta}(b^*), \qquad (2.14)$$

Именно равенства (2.14) чаще всего и используют для практической оценки погрешности.

Итак, выполнение арифметических операций над приближенными числами, как правило, сопровождается потерей точности. Единственная операция, при которой потеря не происходит, это сложение чисел одного знака. Наибольшая потеря точности может произойти при вычитании близких чисел одного знака.

§ 2.4. Погрешность функции

1.Погрешность функции многих переменных. Пусть $f(x) = f(x_1, x_2, ..., x_m)$ – дифференцируемая в области G функция m переменных, вычисление которой производится при приближенно заданных аргументах $x_1^*, x_2^*, \dots, x_m^*$. Такая ситуация возникает, например, всякий раз, когда на ЭВМ производится расчет по формуле. Важно знать, какова величина неустранимой ошибки, вызванной тем, что вместо значения $y = f(x)$ в действительности вычисляется значение $y^* = f(x^*)$, *где* $x^* = (x_1^*, x_2^*, \dots, x_m^*)$.

Введем обозначения: пусть $[x, x^*]$ – отрезок, соединяющий точки x и x^*, и $f'_{x_j} = \partial f / \partial x_j$.

Т е о р е м а 4. Для абсолютной погрешности значения $y^* = f(x^*)$ *справедлива следующая оценка:*

$$\Delta(y^*) \le \sum_{j=1}^{m} \max_{[x,x^*]} \left| f'_{x_j} \right| \Delta(x_j^*). \qquad (2.15)$$

□ Оценка (2.15) вытекает из формулы конечных приращений Лагранжа:

$$f(x) - f(x^*) = \sum_{j=1}^{m} f'_{x_j}(\tilde{x})(x_j - x_j^*), \ \tilde{x} \in [x, x^*].$$

Далее берем модуль от правой и левой частей уравнения и правую часть заменяем на максимум. Получаем требуемое соотношение. ■

С л е д с т в и е. *Если* $x^* \approx x$, *то в силу оценки (2.15) можно положить*

$$\bar{\Delta}(y^*) \approx \sum_{j=1}^{m} |f'_{x_j}(x^*)| \ \bar{\Delta}(x_j^*), \qquad (2.16)$$

$$\bar{\Delta}(y^*) \approx \sum_{j=1}^{m} |f'_{x_j}(x)| \ \bar{\Delta}(x_j^*), \qquad (2.17)$$

Равенство (2.16) удобно для практических оценок, а равенством (2.17) мы воспользуемся в дальнейшем для теоретических построений.

Из формул (2.16) и (2.17) вытекают приближенные равенства для оценки границ относительных погрешностей:

$$\bar{\delta}\,(y^*) \approx \sum_{j=1}^{m} \nu_j^*\ \bar{\delta}\,(x_j^*), \quad \bar{\delta}\,(y^*) \approx \sum_{j=1}^{m} \nu_j\ \bar{\delta}\,(x_j^*). \tag{2.18}$$

Здесь

$$\nu_j^* = \frac{|x_j^*|\cdot|f'_{x_j}(x^*)|}{|f(x^*)|}, \qquad \nu_j = \frac{|x_j|\cdot|f'_{x_j}(x)|}{|f(x)|}, \tag{2.19}$$

2. Погрешность функции одной переменной. Формулы для границ погрешностей функции одной переменной являются частным случаем формул (2.16) – (2.18) при $m = 1$:

$$\bar{\Delta}\,(y^*) \approx |\,f'(x^*)|\ \bar{\Delta}\,(x^*), \quad \bar{\Delta}\,(y^*) \approx |\,f'(x)|\ \bar{\Delta}\,(x^*), \tag{2.20}$$

$$\bar{\delta}(y^*) \approx \nu^*\ \bar{\delta}(x^*), \quad \bar{\delta}(y^*) \approx \nu\ \bar{\delta}(x^*), \tag{2.21}$$

где $\nu^* = |\,x^*|\,|f'(x^*)\,|\,/\,|\,f(x^*)|$, $\nu = |x|\,|f'(x)|\,/\,|\,f(x)|$.