Введение

Численные методы (вычислительные методы, методы вычислений) — раздел вычислительной математики, изучающий приближенные способы решения типовых математических задач, которые либо не решаются, либо трудно решаются точными аналитическими методами (вычислительная математика в узком смысле). Примерами типовых задач являются численное решение уравнений, численные дифференцирование и интегрирование и др. Кроме численных методов, к вычислительной математике относят круг вопросов, связанных с использованием компьютеров и с программированием.

Деление методов вычислений на аналитические и численные несколько условно.

Пример 1. При аналитическом решении квадратного уравнения $ax^2+bx+c=0$ по известной формуле $x_{1,2}=\left(-b\pm\sqrt{b^2-4ac}\right)\Big/(2a)$ в ответ входит корень $\sqrt{\ldots}$. Если он не извлекается точно (подкоренное выражение не является точным квадратом некоторого числа), то для получения численного значения корней потребуется численная процедура приближенного вычисления корня.

Пример 2. Рассмотрим дифференциальное уравнение 1-го порядка с разделяющимися переменными

$$\frac{\mathrm{d}y}{\mathrm{d}x}=y\frac{\sin ax}{x}.$$

Оно легко «решается» аналитически

$$y(x)=C\exp\left\{\int_0^x\frac{\sin at}{t}\,\mathrm{d}t\right\},$$

но интеграл «неберущийся», и вычислять его придется численно.

Итак, даже в тех случаях, когда можно далеко продвинуться в аналитическом решении задачи, не исключено применение на каком-либо этапе численных методов для получения ответа в практически удобном виде.

Часто аналитические методы называют точными, а численные — приближенными. Приведенные примеры показывают, что и аналитические методы могут приводить к приближенному результату. Кроме того, аналитические методы часто бывают приближенными по существу, оставаясь аналитическими, например, когда функция заменяется первыми слагаемыми ее ряда Тейлора.

Элементы теории погрешности

Классификация погрешностей

Поскольку численные методы предназначены для отыскания приближенного решения задач, не решаемых точными методами, такому решению всегда свойственна некоторая погрешность. Рассмотрим здесь источники погрешности.

1) **Погрешность модели.** Природа слишком сложна и многообразна, чтобы пытаться изучать ее во всей полноте присущих ей в том числе и малозначимых взаимосвязей. Любая (естественная) наука изучает не природу непосредственно, а те модели, которые создаются самой этой наукой для описания природных явлений. **Модель** — это идеализированное описание явления, в котором выявлены основные и игнорируются второстепенные свойства явления. Хорошая модель — это верный шарж, меткая карикатура на изучаемое явление. Естественно, что моделирование, сопровождаемое огрублением и упрощением, вносит погрешность в результат описания явления. Математическая модель создается на языке математики, но оценка погрешности математической модели есть прерогатива не математики, а той науки, в рамках которой изучается явление.

2) **Погрешность исходных данных**. Как правило, математическая модель содержит некоторые параметры, зависящие от исходных данных. Поскольку последние определяются обычно из экспериментов, неизбежно сопровождаемых ошибками измерений, возникает погрешность исходных данных.

Погрешности в решении, обусловленные моделированием и исходными данными, называются **неустранимыми**. Они не зависят от математики и присутствуют, даже если решение поставленной математической задачи найдено точно.

3) **Погрешность метода**. После того как математическая модель создана, вычисления в рамках модели обычно можно выполнять поразному. Сложная математическая задача заменяется более простой. Например, вычисление определенного интеграла заменяется вычислением интегральной суммы. При этом неизбежно возникает погрешность метода вычислений, которой в дальнейшем мы будем уделять большое внимание при рассмотрении конкретных численных методов.

4) **Погрешность округления**. Любые расчеты, выполняемые как вручную, так и с помощью вычислительной техники, производятся с конечным числом цифр, поэтому приходится прибегать к округлению промежуточных и окончательного ответа. Так возникает погрешность округления, которая может накапливаться в ходе вычислений (опасный процесс, способный обесценить результат вычислений!). Даже те результаты, которые получены точными аналитическими методами, испытывают влияние погрешности округлений и в действительности могут оказаться приближенными.

Полная погрешность является результатом взаимодействия разных видов погрешностей и не может быть меньше, чем наибольшая из составляющих ее погрешностей.

Абсолютная и относительная погрешности

Для оценки погрешности вводятся понятия абсолютной и относительной погрешности.

Пусть x — точное значение некоторой величины (нам оно неизвестно и никогда не будет известно, поскольку определяется с помощью измерений, страдающих неточностями); a — приближенное значение той же величины ($a \approx x$). Абсолютная погрешность приближенного числа a определяется как $\Delta_a = |x - a|$. Но поскольку x неизвестно, то и абсолютную погрешность мы узнать не можем! Чтобы разрешить парадокс, вводят предельную абсолютную погрешность Δ_a^* — такое значение, которое абсолютная погрешность заведомо не превзойдет при данном способе измерений

$$|x - a| \le \Delta_a^*. \qquad (1)$$

Из выражения (1) следует, что $a - \Delta_a^* \le x \le a + \Delta_a^*$, поэтому желательно возможно меньшее значение Δ_a^* — это уменьшит длину интервала, содержащего искомое значение x и, следовательно, понизит неопределенность в наших знаниях об этой величине.

В технике формулу (1) часто записывают в виде $x = a \pm \Delta_a^*$, причем Δ_a^* называется *допусками*. Никакое изделие не может быть изготовлено с абсолютно точным соблюдением номинальных размеров, допуски показывают возможные (*допустимые*) отклонения от номинала.

Итак, абсолютная погрешность оценивает точность измерений, но эта оценка *неполна*, поскольку не учитывает характерный размер изучаемого явления (объекта). Так, например, абсолютная погрешность в 1 см при измерении длины комнаты — вероятно, вполне приемлемая точность, но при измерении роста человека эта же погрешность будет сочтена непозволительно грубой.

Более информативным показателем качества измерений является *относительная погрешность* δ_a (соответственно *предельная относительная погрешность* δ_a^*) приближенного числа a как отношение абсолютной погрешности (предельной абсолютной погрешности) к модулю числа a

$$\delta_a = \frac{\Delta_a}{|a|}, \quad \delta_a^* = \frac{\Delta_a^*}{|a|}.$$

Относительная погрешность является величиной *безразмерной*, т. е. не зависит от выбора системы единиц измерения, что позволяет сравнивать качество измерений разнородных величин (бессмысленным является вопрос о том, что больше: 1 кг или 1 м, — но сравнение *качества* измерений массы и длины в терминах относительной погрешности вполне допустимо). Измеряется δ_a (δ_a^*) в долях единицы или в процентах.

Пример. Согласно ныне действующим (2015 г.) определениям международного Комитета по константам для науки и технологии входящая в закон всемирного тяготения гравитационная постоянная

$$\gamma = (6.67259 \pm 0.00085) \cdot 10^{-11}\ \text{м}^3 \cdot \text{кг}^{-1} \cdot \text{с}^{-2},$$

а заряд электрона

$$e = (1.602\,177\,33 \pm 0.000\,000\,49) \cdot 10^{-19}\ \text{Кл}.$$

Сравнить точность определения этих фундаментальных физических постоянных.

Решение. Для гравитационной постоянной предельная относительная погрешность

$$\delta_{\gamma}^{*} = \frac{0.00085}{6.67259} = 1.27 \cdot 10^{-4},$$

а для заряда электрона

$$\delta_{e}^{*} = \frac{0.000\,000\,49}{1.602\,177\,33} = 3.1 \cdot 10^{-7}.$$

Таким образом, в последнем случае относительная погрешность оказывается на три порядка меньшей, т. е. заряд электрона определен существенно точнее, чем гравитационная постоянная.

С понятиями абсолютной и относительной погрешности связаны понятия верных и значащих цифр.

Если абсолютная погрешность приближенного числа не превышает единицы последнего (самого правого) разряда его десятичной записи, то цифры числа называют **верными** (или **точными**).

По умолчанию десятичная запись приближенного числа должна содержать только верные цифры, и тогда по записи числа сразу можно узнать предельную абсолютную погрешность, с которой оно известно.

Цифры, не являющиеся верными, называются **сомнительными**.

Пример. Даны приближенные числа $a = 8.6$, $b = 8.60$, $c = 3200$, $d = 3.2 \cdot 10^{3}$. Указать предельную абсолютную погрешность для каждого числа.

Решение. Для числа a погрешность $\Delta_{a}^{*} \le 0.1$, для числа b $\Delta_{b}^{*} \le 0.01$, для числа c $\Delta_{c}^{*} \le 1$, для числа d $\Delta_{d}^{*} \le 0.1 \cdot 10^{3} = 100$.

Итак, числа a и b, c и d, равные с точки зрения «обычной» математики, существенно различны в вычислительной математике: из абсолютной погрешности мы заключаем, что число b известно точнее, чем число a, а число c — точнее, чем d. Кроме того, нуль, стоящий справа в дробной части десятичного числа, важен, и им нельзя пренебрегать, если мы хотим составить верное суждение о точности числа.

Значащими цифрами приближенного числа называются все цифры его десятичной записи, кроме нулей, находящихся левее первой отличной от нуля цифры.

Пример. Числа 0.001 307 и 6.0400 имеют соответственно четыре и пять значащих цифр. Итак, нули, находящиеся слева, значащими не являются, а нуль, записанный в конце десятичной дроби, всегда является значащей цифрой.

Действия с приближенными числами

Теорема 1. Абсолютная погрешность алгебраической суммы нескольких приближенных чисел не превышает суммы абсолютных погрешностей этих чисел.

В частности, для суммы двух чисел a и b любого знака получаем $\Delta_{a\pm b} \le \Delta_a + \Delta_b$.

Из этой теоремы следует, что абсолютная погрешность алгебраической суммы не меньше абсолютной погрешности наименее точного из слагаемых, т. е. увеличение точности за счет других слагаемых невозможно. Поэтому бессмысленно сохранять излишние десятичные знаки в более точных слагаемых. Отсюда вытекает следующее.

Правило сложения и вычитания приближенных чисел:

1) выделить наименее точное число (или числа), т. е. такое, в десятичной записи которого наименьшее число верных десятичных знаков;

2) округлить остальные числа так, чтобы каждое из них содержало на один (запасной) знак больше, чем выделенное число;

3) выполнить сложение и вычитание с учетом сохраненных знаков;

4) полученный результат округлить до предпоследнего знака.

Напомним правила округления числа, т. е. его замены числом с меньшим количеством значащих цифр:

1) если первая из отбрасываемых цифр меньше 5, то сохраняемые десятичные знаки оставляют без изменения;

2) если первая из отбрасываемых цифр больше 5, то последний из сохраняемых знаков увеличивают на 1;

3) если первая из отбрасываемых цифр равна 5, а среди следующих за ней цифр есть отличные от нуля, то последний из сохраняемых знаков увеличивают на 1;

4) если первая из отбрасываемых цифр равна 5, а все последующие — нули, то последний из сохраняемых десятичных знаков увеличивают

на 1, когда он нечетен, и сохраняют неизменным, когда он четен (правило четной цифры).

Пример. Округляя число 53.471 до одного знака после запятой, получим 53.5 (правило 2), а при округлении до двух знаков после запятой получим 53.47 (правило 1). Округляя число 7.825 001 до трех знаков после запятой, получим 7.825 (правило 3). Округляя число 8.465 до сотых долей, получим 8.4**6**; сохраняемая цифра не увеличивается на единицу, поскольку она четна. При округлении числа 8.475 до сотых долей получим 8.4**8** — нечетная цифра увеличилась на единицу (правило 4).

Смысл правила 4 в том, что при многочисленных округлениях избыточные числа будут встречаться примерно с той же частотой, что и недостаточные, и произойдет частичная взаимная компенсация погрешностей округления; результат окажется более точным.

Теперь проиллюстрируем правило сложения и вычитания приближенных чисел.

Пример. Найти сумму приближенных чисел $a = 414.8$, $b = 0.025$, $c = 24.17$, $d = 0.000\,326$. По умолчанию все цифры в этих числах считать верными.

Решение. Наименее точное слагаемое — a, поскольку в нем только один верный десятичный знак. Округлим остальные слагаемые до двух знаков после запятой: $b \to 0.02$, $c \to 24.17$, $d \to 0.00$. Теперь сложим округленные числа: $414.8 + 0.02 + 24.17 + 0.00 = 438.99$. Округляя результат до одного знака после запятой, получим окончательный ответ: 439.0.

Теорема 2. Относительная погрешность произведения (частного) приближенных чисел не превышает суммы относительных погрешностей этих чисел.

В частности, для трех чисел $\delta_{ab/c} \le \delta_a + \delta_b + \delta_c$.

Из теоремы следует, что относительная погрешность произведения и частного не может быть меньше относительной погрешности наименее точного из исходных чисел (т. е. имеющего меньше всего верных значащих цифр). Поскольку относительная погрешность числа определяется количеством его верных значащих цифр, то при умножении и делении бессмысленно оставлять значащих цифр больше, чем их было в исходном числе с наименьшим количеством верных значащих цифр.

Отсюда вытекает следующее правило.

Правило умножения и деления приближенных чисел:

1) из всех чисел, которые предстоит умножать и делить, выделить наименее точное — то, в котором меньше всего верных значащих цифр;

2) округлить остальные числа так, чтобы каждое из них содержало на одну (з а п а с н у ю) значащую цифру больше, чем выделенное число;

3) выполнить умножение и деление округленных чисел с учетом сохраненных значащих цифр;

4) оставить в ответе столько значащих цифр, сколько их было в наименее точном числе.

Пример. Найти произведение приближенных чисел $a = 3.5$ и $b = 83.368$, все цифры которых верные.

Решение. В первом числе две верные значащие цифры, а во втором — пять. Второе число округлим до трех значащих цифр: $b \to 83.4$. После округления перемножим числа: $ab = 3.5 \cdot 83.4 = 291.9 \approx 2.9 \cdot 10^2$. В ответе оставлены две значащие цифры — столько, сколько их было во множителе с наименьшим количеством верных значащих цифр.

Численное решение нелинейных уравнений

Корнем уравнения $f(x)=0$ называется значение $x=\bar{x}$, подстановка которого в уравнение превращает его в верное числовое равенство. Например, если в уравнение $x^2+5x+4=0$ подставить $x=-1$, то получим $0=0$ (верно). Решить уравнение — значит найти его корни. Далеко не каждое уравнение допускает аналитическое решение:

1) трансцендентные уравнения, как правило, не решаются аналитически, за исключением специальных случаев («школьного» типа), когда уравнение можно удачной подстановкой свести к алгебраическому, например, $e^{2x}-6e^x+9=0$;

2) даже для алгебраического уравнения

$$a_0x^n+a_1x^{n-1}+\ldots+a_{n-1}x+a_n=0$$

степени выше четвертой не существует формулы, выражающей корни через коэффициенты уравнения при помощи конечного числа арифметических операций и извлечения корней (в частных случаях, например для уравнения $x^{42}-5x^{21}+4=0$, такие формулы могут существовать, но в общем случае нет). Невозможность аналитического решения уравнений степени пятой и высших доказана трудами Абеля (1802–1829) и Галуа (1811–1832).

Таким образом, большое значение имеет задача приближенного, численного отыскания корней уравнений, для этого:

а) определяют количество корней уравнения и изолируют (отделяют) каждый из них. **Отрезком изоляции** называется отрезок, на котором лежит только один корень уравнения;

б) вычисляют каждый корень с требуемой точностью.

Для отделения корней уравнения $f(x)=0$ применяют графический и аналитический методы.

В первом из них строят график функции $y=f(x)$ и приближенно находят точки его пересечения с осью Ox.

Пример. Для отделения корней уравнения $x^2 - 4x + 5 = 0$ строим график функции $f(x) = x^2 - 4x + 5$ (рис. 8). График пересекает ось абсцисс в единственной точке на отрезке $[-3, -2]$, который и будет отрезком изоляции корня.

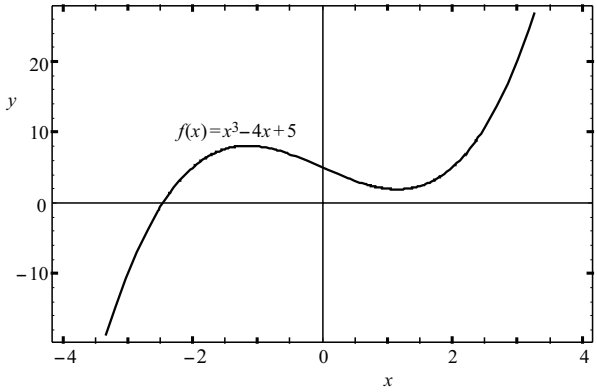

Рис. 8. Графическое отделение корней
(единственный корень уравнения $x^2 - 4x + 5 = 0$ лежит на отрезке $[-3, -2]$)

Аналитический способ отделения корней уравнения $f(x) = 0$ основан на том, что для функции $f(x)$, непрерывной на отрезке $[a, b]$ и принимающей на его концах значения разных знаков, существует по меньшей мере одна точка $\bar{x} \in [a, b]$, такая, что $f(\bar{x}) = 0$. Если на этом отрезке функция $f(x)$ монотонна, то корень $\bar{x}$ единственный, в противном случае корней может быть несколько (рис. 9).

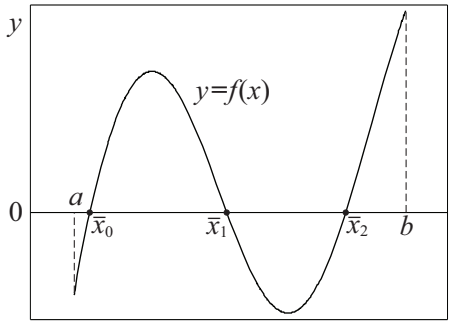

Рис. 9. Отделение корней в случае немонотонности функции[7]

[7] На концах отрезка $[a, b]$ функция $f(x)$ принимает значения разных знаков; поскольку она немонотонна на этом отрезке, то уравнение $f(x) = 0$ имеет несколько корней (точки $\bar{x}_0, \bar{x}_1, \bar{x}_2$).

Начнем теперь рассмотрение методов вычисления корней с заданной точностью.

Метод половинного деления (дихотомия[8])

Метод непосредственно следует из аналитического способа отделения корней. Пусть для уравнения $f(x)=0$ найден первичный отрезок $[x_0, x_1]$ изоляции корня. Вычислим середину отрезка $x_2 = \frac{x_0 + x_1}{2}$. Если случайно окажется, что $f(x_2)=0$, то x_2 является корнем уравнения $f(x)=0$. Если же $f(x_2)\neq 0$, то из двух половин $[x_0, x_2]$, $[x_2, x_1]$ первичного отрезка выберем для дальнейшего деления пополам ту, на концах которой функция $f(x)$ принимает значения противоположных знаков. Выбранный отрезок снова разделим пополам и найдем половину с противоположными знаками $f(x)$ на концах, и т. д.

Критерий достижения требуемой точности (критерий обрыва счета): если корень надо вычислить с точностью ε, то деление пополам следует продолжать до тех пор, пока длина очередного отрезка не станет меньше 2ε; тогда середина этого отрезка даст значение корня с точностью ε.

Свойства дихотомии следующие:

а) идейная простота метода;

б) непритязательность к свойствам функции $f(x)$ — она должна быть лишь непрерывной, а дифференцируемость не предполагается.

К сожалению, отрицательные свойства перевешивают:

в) очень медленная сходимость. Пусть, например, первичный отрезок изоляции имеет единичную длину. После первого шага дихотомии длина уменьшится до $1/2$, после второго — до $(1/2)^2$, и т. д. Поскольку $(1/2)^{10} = 1/1024 < 0{,}001$, после десяти шагов дихотомии обеспечиваются лишь три верных десятичных знака искомого корня[9];

г) неприменимость к вычислению корней четной кратности (рис. 10);

д) неприменимость к решению систем уравнений.

---

[8] Дихотомия (греч. Διχοτομία) — разрубание пополам, разделение надвое.

[9] Гора родила мышь.

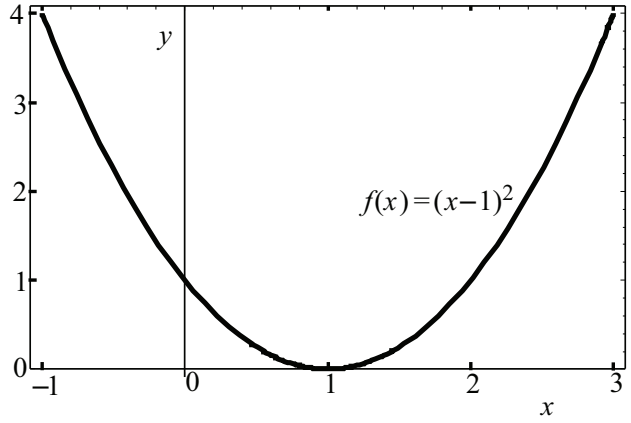

Рис. 10. Отделение корней: случай кратного корня[10]

Метод итераций (последовательных приближений)

Пусть имеется уравнение $f(x)=0$. Приведем его к равносильному виду $x=\varphi(x)$, удобному для итераций[11] (ниже покажем, как это сделать).

Выберем некоторое начальное приближение x_0 и найдем следующие приближения, выполняя однообразные вычисления (итерации),

$$x_1=\varphi(x_0),\ x_2=\varphi(x_1),\ldots,\ x_n=\varphi(x_{n-1}),$$

Отсюда понятно удобство для итераций перехода от записи уравнения в виде $f(x)=0$ к виду $x=\varphi(x)$: значение аргумента в левой части равенства является следующим приближением по отношению к тому, которое подставлялось в функцию $\varphi(x)$. При подстановке значения аргумента в $f(x)$ справа от знака равенства появляется число 0 без возможности итераций.

Если последовательность $\{x_n\}$ имеет предел, то итерационный процесс $x_n=\varphi(x_{n-1})$, $n=1,2,\ldots,$ называется сходящимся. Пусть функция $\varphi(x)$ непрерывна. Тогда, переходя к пределу $n\to\infty$ в рекуррентном соотношении $x_n=\varphi(x_{n-1})$, можно перенести знак предельного перехода через знак функции

[10] Уравнение $(x-1)^2=0$ имеет двукратный корень $x=1$, но к нему невозможно подступиться методом дихотомии, т. к. по обе стороны от точки $x=1$ знак функции $f(x)=(x-1)^2$ одинаков.

[11] Итерация (лат. iteratio) — повторение, повторное действие.

$$\lim_{n\to\infty} x_n = \lim_{n\to\infty} \varphi\left(x_{n-1}\right) = \varphi\left(\lim_{n\to\infty} x_{n-1}\right).$$

Следовательно, $\lim\limits_{n\to\infty} x_n = \lim\limits_{n\to\infty} x_{n-1} = \bar{x}$ является корнем уравнения $x = \varphi(x)$. Условие и скорость сходимости итерационного процесса определяются в соответствии со следующей теоремой.

Теорема. Пусть корень $\bar{x}$ уравнения $x = \varphi(x)$, а также последовательные приближения к нему x_0, $x_1 = \varphi(x_0), x_2 = \varphi(x_1), \ldots, x_n = \varphi(x_{n-1})$, принадлежат отрезку изоляции $[a, b]$, на котором

$$\left|\varphi'(x)\right| \le q < 1 \tag{30}$$

(число q будем называть коэффициентом сжатия[12]).

Следовательно:

1) отображение $\varphi(x)$ является сжимающим, и итерационный процесс $x_n = \varphi(x_{n-1})$ сходится к корню $\bar{x}$ уравнения $x = \varphi(x)$;

2) критерий достижения требуемой точности ε заключается в том, что как только для абсолютной погрешности Δ n-го приближения к корню выполнится условие

$$\Delta = \left|\bar{x} - x_n\right| \le \frac{q}{1-q}\left|x_n - x_{n-1}\right| < \varepsilon, \tag{31}$$

счет можно оборвать.

Метод итераций сходится при любом выборе начального приближения x_0, лишь бы оно попадало в отрезок $[a, b]$, где выполняется условие сходимости (30). Благодаря этому метод является самоисправляющимся, т. е. ошибка в вычислениях, не выводящая за пределы области сходимости $[a, b]$, не повлияет на конечный результат, т. к. ошибочное значение можно рассматривать как новое начальное значение x_0. Методы вычислений, обладающие свойством самоисправления, особенно надежны.

Из формулы (30) следует, что в качестве значения коэффициента сжатия q можно взять

$$q = \max_{[a,b]} \left|\varphi'(x)\right|. \tag{32}$$

[12] Название не является общепринятым; часто это число называют коэффициентом Липшица (1832–1903).

Из оценки погрешности (31) следует, что скорость сходимости итерационного процесса к корню $\bar{x}$ особенно велика при коэффициенте сжатия $q \approx 0$. Когда q приближается к единице (со стороны меньших значений), сходимость замедляется. При $q \geq 1$ последовательность приближений x_0, $x_1 = \varphi(x_0), x_2 = \varphi(x_1), \ldots, x_n = \varphi(x_{n-1}),$ расходится, и найти корень уравнения $x = \varphi(x)$ с его помощью невозможно. Итак, наиболее благоприятен для вычислений случай $q \approx 0$, поскольку при нем небольшое количество итераций обеспечивает вычисление корня с высокой точностью.

Рассмотрим, как привести уравнение $f(x) = 0$ к виду $x = \varphi(x)$, удобному для итераций, и как обеспечить благоприятное значение q:

1) прибавляя x к обеим частям уравнения $f(x) = 0$, получим $x = f(x) + x$. Обозначая правую часть как новую функцию $f(x) + x = \varphi(x)$, приводим уравнение к нужному виду $x = \varphi(x)$;

2) если окажется, что на отрезке изоляции корня $[a, b]$ для введенной функции значение $\max\limits_{[a,b]} |\varphi'(x)|$ недостаточно мало́, применим более общий прием введения параметра λ: сначала уравнение $f(x) = 0$ преобразуем к равносильному (при $\lambda \neq 0$) уравнению $\lambda f(x) = 0$, а затем прибавим x в обеих частях $x = \lambda f(x) + x$ и, вводя новую функцию $\varphi(x) = \lambda f(x) + x$, получим удобное для итераций уравнение $x = \varphi(x)$. Поскольку согласно (30)–(32) высокая скорость сходимости обеспечивается при $q = \max\limits_{[a,b]} |\varphi'(x)| \approx 0$, выберем на отрезке изоляции $[a, b]$ некоторую точку (например, середину отрезка) x_0 и потребуем, чтобы в ней $\varphi'(x_0) = \lambda f'(x_0) + 1 = 0$. Отсюда найдем значение параметра $\lambda = -1/f'(x_0)$, обеспечивающее благоприятное q (его значение можно найти графически, построив график функции $y = |\varphi'(x)|$ на отрезке $[a, b]$ изоляции корня);

3) часто приводит к цели простой прием — по-другому выразить x из уравнения $x = \varphi(x)$, если первый вариант оказался неудачным. Смысл этой рекомендации станет ясен из нижеследующего примера.

Пример. Методом итераций найти корни уравнения

$$5x - 6 \ln x - 7 = 0 \tag{33}$$

с точностью $\varepsilon = 10^{-3}$.

Решение. При помощи графического метода найдем количество корней и отрезки их изоляции. Если график строится вручную, то его

построение для функции $f(x)=5x-6\ln x-7$ затруднительно. Проще привести уравнение к виду $\ln x=\dfrac{5x-7}{6}$ и найти абсциссы точек пересечения графиков функций $f_1(x)=\ln x$ и $f_2(x)=\dfrac{5x-7}{6}$ (рис. 11). При работе, например, в пакете MathCad разбиение функции на $f_1(x)$ и $f_2(x)$ излишне!

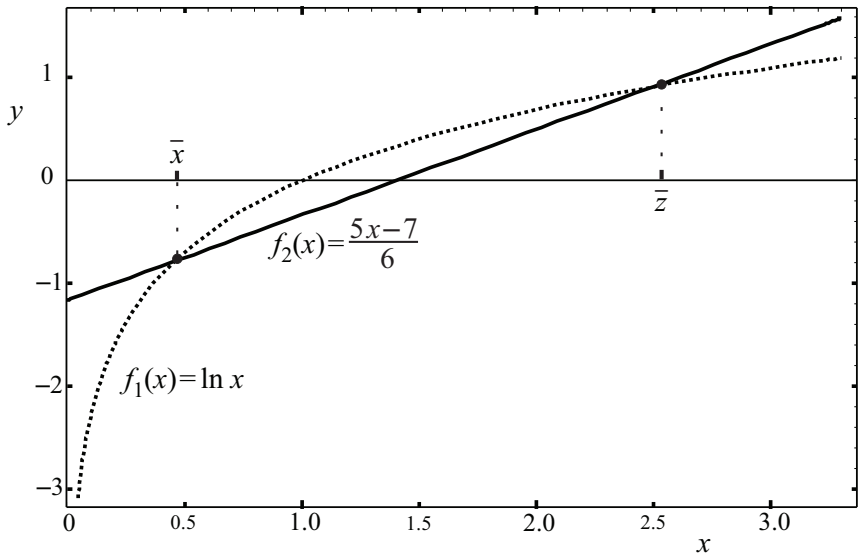

Рис. 11. Графическое отделение корней[13]

Приведем уравнение (33) к виду, удобному для итераций. Можно, например, выразить x из первого слагаемого

$$x=\frac{6\ln x+7}{5},$$

тогда в итерационном процессе будет использоваться функция $\varphi(x)=\dfrac{6\ln x+7}{5}$, и нужно проверить, будет ли такой процесс сходящимся. Для этого вычислим производную $\varphi'(x)=\dfrac{6}{5x}$ и найдем коэф-

[13] Уравнение $5x-6\ln x-7=0$ имеет два корня: $\bar{z}\in[0{,}1;\ 1]$ и $\bar{z}\in[2;\ 3]$. Первый из отрезков изоляции не должен начинаться в нуле, т. к. в этой точке функция $\ln x$ терпит разрыв.

фициент сжатия $q = \max\left|\varphi'(x)\right|$ на каждом отрезке изоляции корня. На правом отрезке изоляции $[2;3]$

$$q_1 = \max_{[2;3]}\left|\varphi'(x)\right| = \max_{[2;3]}\frac{6}{5x} = \left.\frac{6}{5x}\right|_{x=2} = 0.6 < 1.$$

Итерационный процесс будет сходящимся, и его можно использовать для нахождения корня $\bar{z}$.

На левом отрезке $[0.1;1]$ изоляции корня

$$\max_{[0,1;1]}\left|\varphi'(x)\right| = \max_{[0,1;1]}\frac{6}{5x} = \left.\frac{6}{5x}\right|_{x=0,1} = 12 > 1,$$

и функция $\varphi(x) = \dfrac{6\ln x + 7}{5}$ непригодна для вычисления корня $\bar{x}$, т. к. итерационный процесс будет расходящимся. Поэтому для отрезка $[0.1;1]$ по-другому выразим x из уравнения (33), а именно: выразим x, который был аргументом логарифма $x = e^{\frac{5x-7}{6}}$, и введем функцию

$$\psi(x) = \exp\left(\frac{5x-7}{6}\right),$$

которую надо проверить на пригодность к использованию в итерационном процессе. Поскольку

$$q_2 = \max_{[0,1;1]}\left|\psi'(x)\right| = \max_{[0,1;1]}\frac{5}{6}\exp\left(\frac{5x-7}{6}\right) = \frac{5}{6}\exp\left(\frac{5x-7}{6}\right)\Bigg|_{x=1} =$$

$$= \frac{5}{6}e^{-1/3} = \frac{5}{6\sqrt[3]{e}} \approx 0.597 \approx 0.6 < 1,$$

функция $\psi(x)$ обеспечит сходимость итерационного процесса (случайно коэффициент сжатия совпал с q_1).

Резюмируем:

а) для нахождения корня $\bar{x} \in [0.1;1]$ строим итерационный процесс

$$x_{n+1} = \exp\left(\frac{5x_n - 7}{6}\right),\ n = 0, 1, 2, \ldots;$$

начальное приближение $x_0 = 0.5$; коэффициент сжатия $q_2 = 0.6$;

критерий достижения требуемой точности — как только для абсолютной погрешности Δ выполнится условие

$$\Delta = |\bar{x} - x_n| \le \frac{q_2}{1-q_2}|x_n - x_{n-1}| < \varepsilon = 0.001,$$

счет можно оборвать (ответ записать с тремя верными десятичными знаками, гарантируемыми достигнутой точностью);

б) для корня $\bar{z} \in [2;3]$ строим итерационный процесс

$$z_{n+1} = \frac{6\ln z_n + 7}{5},\ n = 0,1,2,\ldots;$$

начальное приближение $z_0 = 2.5$; коэффициент сжатия $q_1 = 0.6$; критерий достижения требуемой точности — как только выполнится условие

$$\Delta = |\bar{z} - z_n| \le \frac{q_1}{1-q_1}|z_n - z_{n-1}| < \varepsilon = 0.001,$$

счет можно оборвать.

В качестве начального приближения к корню традиционно (но не обязательно) берут середину отрезка его изоляции.

Результаты вычислений сведем в таблицы (табл. 3).

Таблица 3

Результаты вычислений

Номер итерации	Приближение к корню	Следующее приближение	Оценка погрешности
n	x_n	$\psi(x_n) = \exp\left(\frac{5x_n - 7}{6}\right)$	$\frac{q_2}{1-q_2}\lvert x_n - x_{n-1}\rvert$
0	0.5	0.4724	–
1	0.4724	0.4616	0.016
2	0.4616	0.4575	$6.177\cdot10^{-3}$
3	0.4575	0.4559	$2.351\cdot10^{-3}$
4	0.4559	0.4553	$8.926\cdot10^{-4} < \varepsilon$
n	z_n	$\varphi(z_n) = \frac{6\ln z_n + 7}{5}$	$\frac{q_1}{1-q_1}\lvert z_n - z_{n-1}\rvert$
0	2.5	2.4995	–
1	2.4995	2.4993	$6.767\cdot10^{-4} < \varepsilon$

На нулевой итерации оценить погрешность невозможно (поставлен прочерк), поскольку никакого значения, предшествующего нулевому приближению, не существует.

Итак, $\bar{x} \approx x_4 = 0.455$; $\bar{z} \approx z_1 = 2.499$.

Свойства метода итераций:

а) дифференцируемость функций, участвующих в расчетах;

б) самоисправляемость вычислительного процесса;

в) скорость сходимости зависит от величины коэффициента сжатия q. Благоприятных (близких к нулю) значений q всегда можно достичь введением параметра λ для ускорения сходимости;

г) когда уравнение имеет несколько корней, как правило, для нахождения каждого из них приходится индивидуально строить итерационный процесс, поскольку сходимость одного процесса на разных отрезках изоляции обычно не достигается.

Метод Ньютона

Пусть в уравнении $f(x) = 0$ функция $f(x)$ имеет непрерывную производную $f'(x) \neq 0$; x_n есть некоторое приближение к корню $\bar{x}$ рассматриваемого уравнения. В окрестности точки x_n разложим функцию $f(x)$ в ряд Тейлора

$$f(x) = f(x_n) + f'(x_n)(x - x_n) + \frac{f''(x_n)}{2!}(x - x_n)^2 + \\ + \frac{f'''(x_n)}{3!}(x - x_n)^3 + \ldots + \frac{f^{(k)}(x_n)}{k!}(x - x_n)^k + \ldots$$

и ограничимся линейным по x слагаемым включительно

$$0 = f(x) \approx f(x_n) + f'(x_n)(x - x_n). \qquad (34)$$

Отсюда

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

и согласно идее Ньютона левую часть этого выражения будем рассматривать как следующее, $(n+1)$-е, приближение некоторого итерационного процесса

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad n = 0, 1, 2, \ldots \qquad (35)$$

Формула (35) представляет метод Ньютона численного решения уравнений. Другое название — метод линеаризации, поскольку функция $f(x)$ приближенно заменена линейной (34).

Выясним геометрический смысл итерационного процесса (35). В точке с абсциссой x_0 проведем касательную к графику функции $y = f(x)$ (рис. 12); уравнение касательной

$$y - f(x_0) = f'(x_0)(x - x_0).$$

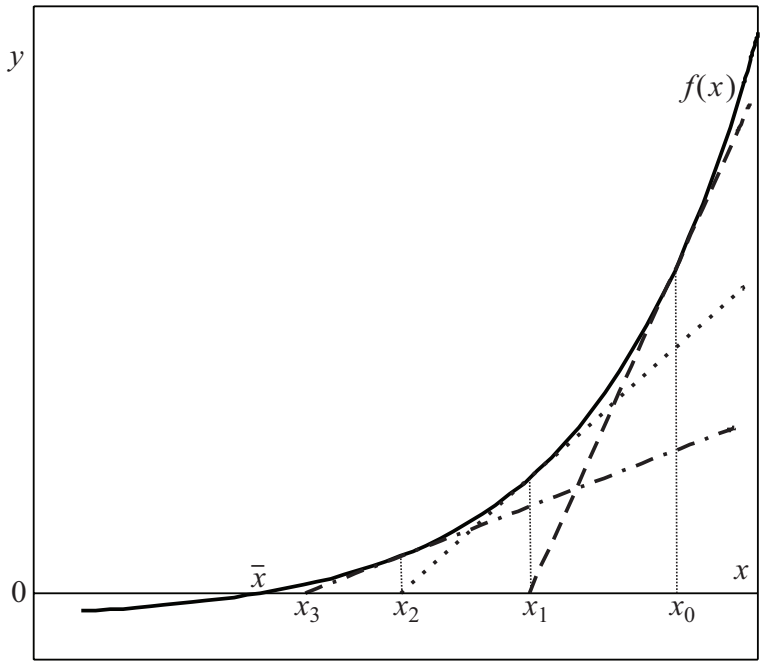

Рис. 12. Геометрическая иллюстрация метода Ньютона

Найдем точку пересечения данной прямой с осью абсцисс (в этой точке $y = 0$)

$$x = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

Согласно формуле (35) полученное значение — это следующее приближение x_1. В точке с абсциссой x_1 проведем еще одну касательную к графику $f(x)$; уравнение касательной

$$y - f(x_1) = f'(x_1)(x - x_1).$$

Точка пересечения данной прямой с осью абсцисс

$$x = x_1 - \frac{f(x_1)}{f'(x_1)}.$$

Это будет второе приближение x_2, и т. д.

Итак, на каждой итерации график функции $f(x)$ заменяется его касательной. Поэтому метод Ньютона называют еще методом касательных.

Метод Ньютона можно рассматривать как частный случай рассмотренного выше метода итераций. В самом деле, от уравнения $f(x) = 0$ можно перейти к равносильному

$$-\frac{f(x)}{f'(x)} = 0, \text{ или } x - \frac{f(x)}{f'(x)} = 0 + x.$$

Вводя функцию

$$\varphi(x) = x - \frac{f(x)}{f'(x)},$$

приходим к виду $x = \varphi(x)$, удобному для итераций. Скорость сходимости итерационного процесса, как известно, определяется значением $|\varphi'(x)|$ на отрезке изоляции корня. В нашем случае

$$\varphi'(x) = 1 - \frac{\left[f'(x)\right]^2 - f(x)f''(x)}{\left[f'(x)\right]^2}.$$

Если подставить сюда корень $x = \bar{x}$, то, с учетом равенства $f(\bar{x}) = 0$ (в уравнение подставлен его собственный корень!), получаем

$$\varphi'(\bar{x}) = 0.$$

Таким образом, в точках, очень близких к корню $\bar{x}$ уравнения $f(x) = 0$, скорость сходимости итерационного процесса бесконечно велика! В результате можно сделать вывод о том, что при выборе начального приближения x_0 достаточно близко к $\bar{x}$ метод Ньютона (35) должен обеспечивать быструю сходимость последовательности приближений $x_0, x_1, \ldots, x_k, \ldots$ к искомому корню $\bar{x}$.

К сожалению, этот вывод несколько поспешен. Рассмотрим уравнение $\operatorname{arctg} x = 0$ (рис. 13). Начальное приближение x_0 близко к корню $\bar{x} = 0$, но каждое следующее приближение все дальше от него.

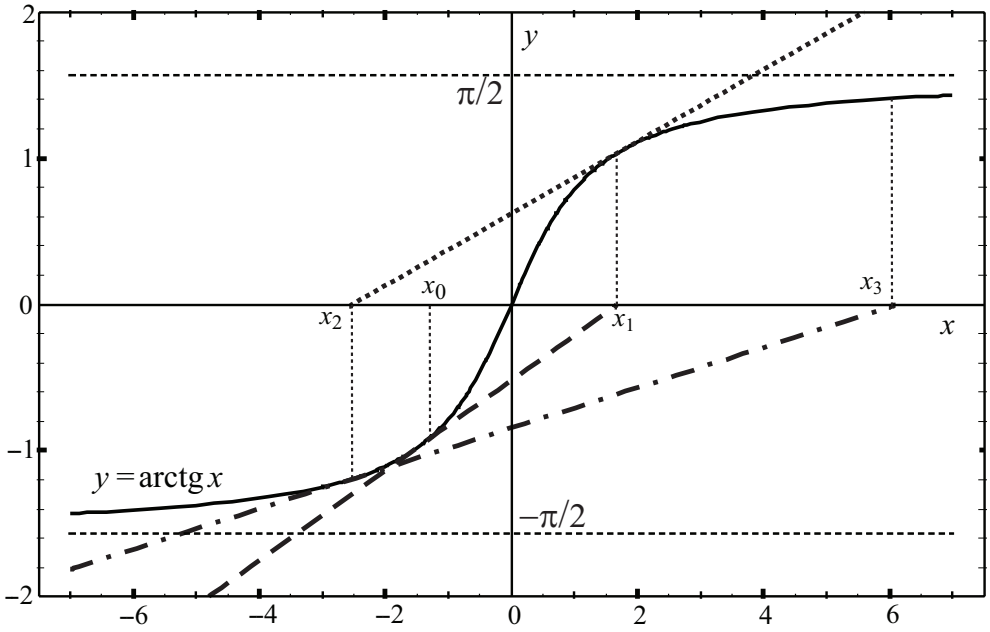

Рис. 13. Пример расходящейся последовательности приближений $x_0, x_1, x_2, x_3, \ldots$ в методе Ньютона

Оказывается, что при сохранении знака производных $f'(x)$, $f''(x)$ на отрезке $[a, b]$ изоляции корня уравнения $f(x) = 0$ метод Ньютона сходится, если в качестве начального приближения x_0 взять любую точку отрезка $[a, b]$.

Если же $f''(x)$ меняет знак на отрезке изоляции корня, то сходимость итерационного процесса не гарантируется. В рассмотренном примере ситуация именно такова: при переходе через $x = 0$ вторая производная

$$(\operatorname{arctg} x)'' = \left(\frac{1}{1+x^2}\right)' = -\frac{2x}{\left(1+x^2\right)^2}$$

меняет знак (направление выпуклости графика арктангенса меняется на противоположное; $x = 0$ — точка перегиба).

Пример. Решить методом Ньютона уравнение

$$5x - 6\ln x - 7 = 0.$$

Решение. Выше (см. рис. 11) корни этого уравнения уже изолированы $\bar{x} \in [0.1; 1]$ и $\bar{z} \in [2; 3]$. Для функции

$$f(x) = 5x - 6\ln x - 7$$

первая и вторая производные $f'(x) = 5 - 6/x$, $f''(x) = 6/x^2$ сохраняют знак на обоих отрезках изоляции корней, что является гарантией сходимости итерационных процессов.

Для левого корня $\bar{x} \in [0.1; 1]$ выбираем начальное приближение, например $x_0 = 0.5$ (в силу самоисправляемости метода это может быть любая точка отрезка изоляции); следующие приближения по Ньютону

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{5x_n - 6\ln x_n - 7}{5 - \dfrac{6}{x_n}} = \frac{6x_n \ln x_n + x_n}{5x_n - 6}, \quad n = \mathbf{0}, 1, 2, \ldots$$

Для правого корня $\bar{z} \in [2; 3]$ при начальном приближении, например $z_0 = 3.5$, итерационный процесс строится точно так же, как для левого,

$$z_{n+1} = z_n - \frac{f(z_n)}{f'(z_n)} = \frac{6z_n \ln z_n + z_n}{5z_n - 6}, \quad n = 0, 1, 2, \ldots$$

Когда нужная точность будет достигнута? В соответствии с известной формулой конечных приращений Лагранжа для дифференцируемой функции $f(x)$

$$f(\alpha) - f(\beta) = f'(\xi) \cdot (\alpha - \beta),$$

где (заранее неизвестная) промежуточная точка $\xi \in [a, b]$. Пусть здесь α равняется x_n, β равняется $\bar{x}$ — искомому значению корня, тогда $f(\beta) = 0$ и

$$f(x_n) = f'(\xi) \cdot (x_n - \bar{x}).$$

Отсюда получаем критерий обрыва счета в методе Ньютона: как только абсолютная погрешность n-го приближения

$$\Delta = |x_n - \bar{x}| = \left| \frac{f(x_n)}{f'(\xi)} \right| \le \frac{|f(x_n)|}{m},$$

где $m = \min\limits_{[a,b]} |f'(x)|$, станет меньше требуемой точности ε

$$\frac{|f(x_n)|}{m} < \varepsilon, \tag{36}$$

точность достигнута, и вычисления можно прекратить, записав ответ $\bar{x} \approx x_n$.

Свойства метода Ньютона таковы:

а) функции, участвующие в расчетах, должны быть дифференцируемыми;

б) вычислительный процесс (35) самоисправляющийся;

в) нахождение всех корней, сколько бы их не было, обслуживается одним и тем же вычислительным процессом (35) — в противоположность методу итераций, в котором, как правило, для каждого корня приходится индивидуально строить итерационный процесс;

г) скорость сходимости итерационного процесса высока;

д) на каждом шаге вычислений требуется вычислять производную $f'(x_n)$, что может иногда представлять проблему при сложно заданной функции.

На частичное устранение этого единственного возможного недостатка метода Ньютона направлено введение двух методов, являющихся его следствиями, — метода секущих и метода хорд.

Метод секущих

Поскольку математически производная вводится как предел отношения приращения функции к приращению ее аргумента

$$f'(x) = \lim_{\Delta x \to 0} \frac{\Delta f(x)}{\Delta x} = \lim_{z \to x} \frac{f(z) - f(x)}{z - x},$$

то, убирая предельный переход, получим приближенное значение производной

$$f'(x) \approx \frac{f(z) - f(x)}{z - x}. \tag{37}$$

В соответствии с этим в методе секущих производная приближенно вычисляется по формуле

$$f'(x_n) \approx \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}. \tag{38}$$

Подставляя в формулу (35) это выражение для производной, приходим к следующему итерационному процессу

$$x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}\ f(x_n), \quad n = 1, 2, 3, \ldots, \tag{39}$$

для которого надо указать д в а начальных приближения x_0 и x_1 (в отличие от метода Ньютона, в котором требовалось указать только x_0). С геометрической точки зрения в методе секущих (39) к а с а т е л ь н а я заменяется с е к у щ е й, проходящей через точки кривой $y = f(x)$ с абсциссами x_n и x_{n-1} (рис. 14).

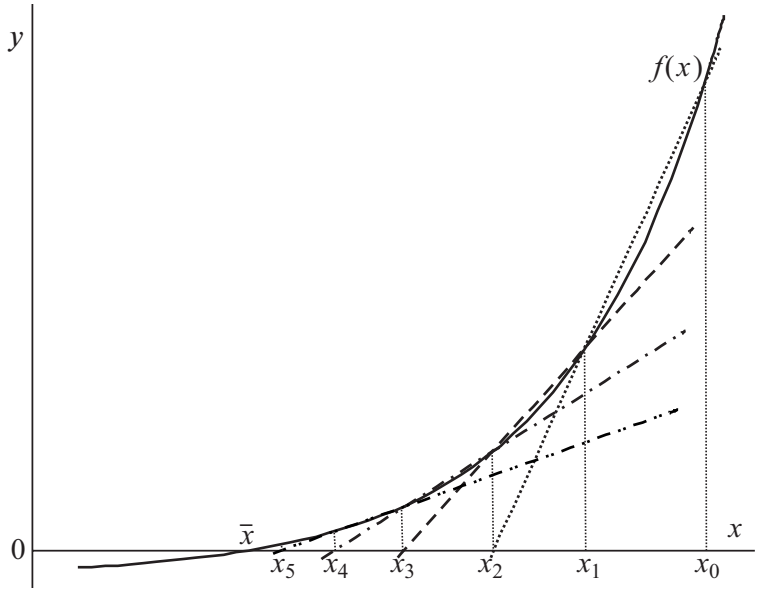

Рис. 14. Геометрическая иллюстрация метода секущих

Свойства метода секущих заключаются в следующем:

а) поскольку метод секущих является модификацией метода Ньютона (метода касательных), он наследует все свойства последнего;

б) скорость сходимости итерационного процесса ниже, чем в методе касательных (вследствие огрубления, заложенного приближением (38), но остается высокой;

в) производная $f'(x_n)$ изгоняется из вычислительного процесса, но лишь частично, поскольку для контроля точности (36) она по-прежнему нужна!

Метод хорд

Применим в формуле (35) еще более грубое приближение для производной

$$f'(x_n) \approx \frac{f(x_n) - f(x_0)}{x_n - x_0}$$

(в самом деле, точки x_n и x_0 дальше друг от друга, чем x_n и x_{n-1} в выражении (38)). Подставляя в формулу (35) это выражение для производной, получим итерационный процесс метода хорд:

$$x_{n+1} = x_n - \frac{x_n - x_0}{f(x_n) - f(x_0)} \; f(x_n), \quad n = 1, 2, 3, \ldots,$$

для которого, как и в методе секущих, надо указать два начальных приближения x_0 и x_1. Геометрически в методе хорд касательная заменяется хордой, проходящей через точки кривой $y = f(x)$ с абсциссами x_n и x_0, (рис. 15).

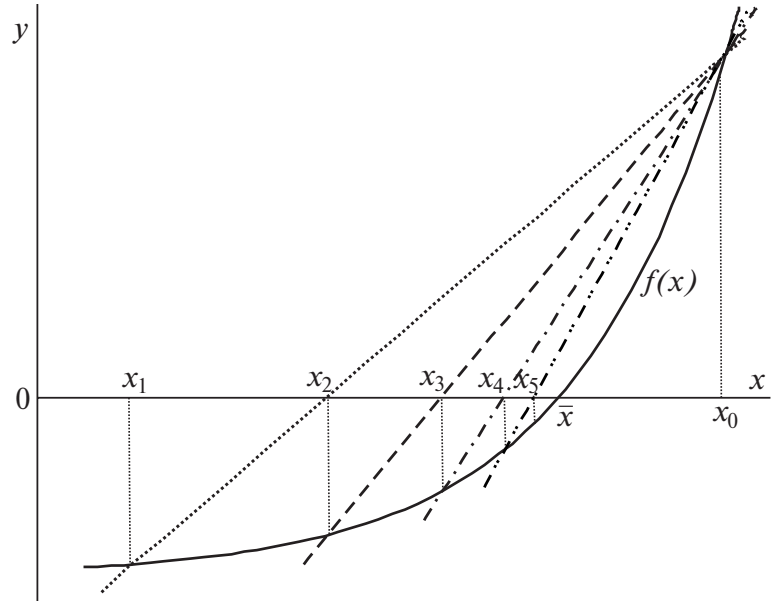

Рис. 15. Геометрическая иллюстрация метода хорд

Метод хорд имеет те же свойства, что и метод секущих, но скорость сходимости при прочих равных условиях становится еще несколько ниже, чем у последнего (оставаясь высокой).

Методы решения систем линейных алгебраических уравнений

Линейные системы

Пусть задана система m линейных алгебраических уравнений с n неизвестными

$$\left.\begin{array}{rcl} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n & = & b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n & = & b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & & \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n & = & b_m. \end{array}\right\} \qquad (40)$$

Вводя матрицу коэффициентов при неизвестных (матрицу системы)

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ . & . & . & . \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix},$$

столбец правых частей (неоднородностей) и столбец неизвестных[14]

$$\boldsymbol{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}, \quad \boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix},$$

можно записать систему (40) в матричном виде

$$A\boldsymbol{x} = \boldsymbol{b}.$$

[14] Эти столбцы формально можно считать векторами.

Методы решения линейных систем можно разбить на две группы: точные (прямые) и приближенные (итерационные).

К точным методам относятся такие, которые в предположении, что вычисления ведутся точно (без округлений), за конечное, заранее оцениваемое количество шагов вычислений приводят к точным значениям неизвестных x_i. Фактически, из-за почти неизбежных округлений при вычислениях, результаты, получаемые точными методами, будут содержать погрешности. Точными являются, например, известные методы Крамера (1704–1752) и Гаусса (1777–1855).

К приближенным относятся такие методы, которые даже в предположении отсутствия погрешности округлений доставляют решение системы лишь с заданной точностью. Точное решение системы достигается асимптотически как результат бесконечного процесса. Примерами приближенных методов являются метод простой итерации [15] и его модификация — метод Зейделя (1821–1896).

Прежде чем переходить к рассмотрению приближенных методов, напомним некоторые особенности методов Крамера и Гаусса.

Правило Крамера применимо при условиях:

а) количество неизвестных n в системе равно числу уравнений m, тогда матрица системы A квадратная, и ей можно сопоставить определитель $\det A$;

б) матрица A невырожденная, т.е. $\det A \neq 0$.

При выполнении этих (довольно стеснительных!) условий решение системы (40) можно найти по формулам Крамера

$$x_i = \frac{\det A_i}{\det A}, \tag{41}$$

где A_i – матрица, получаемая из исходной матрицы A заменой ее i-го столбца столбцом правых частей b_i.

Итак, для решения системы из n уравнений с n неизвестными по правилу Крамера нужно вычислить $(n+1)$ определитель n-го порядка, что очень трудоемко (при самой экономичной организации вычислений потребуется выполнить порядка $\frac{2}{3}n^4$ арифметических операций). Таким образом, правило Крамера удобно

[15] Смысл слова *простая* выяснится ниже, при обсуждении метода Зейделя как модификации данного метода.

в теоретических приложениях (доказывание теорем, вывод формул и т. п.) благодаря компактности формулы (41), но крайне неудобно для практического решения систем.

Почти во всех отношениях более предпочтительным является метод Гаусса (метод последовательного исключения неизвестных). Он не требует невырожденности матрицы A; более того, она может даже не быть квадратной. Если же решается система из n уравнений с n неизвестными, то для реализации метода Гаусса потребуется примерно

$$\frac{2n(n+1)(n+2)}{3}+n(n-1),$$

или порядка $\frac{2}{3}n^3$ арифметических операций (сравните с оценкой для правила Крамера!).

Метод простой итерации

Пусть система из n уравнений с n неизвестными

$$\left.\begin{array}{rcl} a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n &=& b_1,\\ a_{21}x_1+a_{22}x_2+\ldots+a_{2n}x_n &=& b_2,\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & & \\ a_{n1}x_1+a_{n2}x_2+\ldots+a_{nn}x_n &=& b_n \end{array}\right\} \qquad (42)$$

записана в векторно-матричной форме

$$\boldsymbol{Ax}=\boldsymbol{b} \qquad (43)$$

и приведена к виду

$$\boldsymbol{x}=C\boldsymbol{x}+\boldsymbol{f}\ , \qquad (44)$$

удобному для итерирования ($C_{n\times n}$ — некоторая матрица; $\boldsymbol{f}$ — вектор-столбец). Удобство состоит в том, что, задав начальное приближение

$$\boldsymbol{x}^{(0)}=\begin{pmatrix} x_1^{(0)}\\ x_2^{(0)}\\ \vdots\\ x_n^{(0)} \end{pmatrix}$$

и подставляя его в правую часть уравнения (44), слева получим следующее приближение:

$$\boldsymbol{x}^{(1)} = C\boldsymbol{x}^{(0)} + \boldsymbol{f}$$

и

$$\boldsymbol{x}^{(k+1)} = C\boldsymbol{x}^{(k)} + \boldsymbol{f}, \quad k = 0,1,2,\ldots \tag{45}$$

В развернутой записи —

$$\left.\begin{array}{rcl} x_1^{(k+1)} & = & c_{11}x_1^{(k)} + c_{12}x_2^{(k)} + \ldots + c_{1n}x_n^{(k)} + f_1, \\ \ldots & \ldots & \ldots\ldots\ldots\ldots\ldots\ldots \\ x_n^{(k+1)} & = & c_{n1}x_1^{(k)} + c_{n2}x_2^{(k)} + \ldots + c_{nn}x_n^{(k)} + f_n. \end{array}\right\} \tag{46}$$

Выполняя итерации, получим последовательность векторов $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)},\ldots,\boldsymbol{x}^{(k)},\ldots$ Сходится ли эта последовательность, и если да — сходится ли она к искомому решению системы (42)?

Условия сходимости итерационного процесса

Прежде чем сформулировать условия сходимости, напомним некоторые понятия функционального анализа.

Обобщением понятия длины m-мерного вектора $\boldsymbol{x}$ является норма вектора $\|\boldsymbol{x}\|$, из множества способов введения которой в численных методах наиболее употребительными являются следующие три нормы:

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{m}|x_i|, \quad \|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^{m}|x_i|^2}, \quad \|\boldsymbol{x}\|_\infty = \max_{1\le i\le m}|x_i|. \tag{47}$$

Первые две являются частными случаями (при $p=1$ и $p=2$) общей нормы Минковского (1864–1909)

$$\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{m}|x_i|^p\right)^{1/p}, \ p \ge 1,$$

а последняя получается из нее предельным переходом при $p \to \infty$. Заметим, что норма $\|\boldsymbol{x}\|_2$ является естественным обобщением длины вектора, вычисляемой по теореме Пифагора в 2- и 3-мерном геометрических пространствах, на случай m-мерного пространства. Норма $\|\boldsymbol{x}\|_2$ называется евклидовой.

Число

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} \tag{48}$$

называется нормой матрицы A. Поскольку умножение матрицы A на вектор $\boldsymbol{x}$ можно рассматривать как преобразование, переводящее вектор $\boldsymbol{x}$ в новый вектор $\boldsymbol{y} = A\boldsymbol{x}$, то дробь $\|A\boldsymbol{x}\|/\|\boldsymbol{x}\|$ в формуле (48) является не чем иным, как коэффициентом сжатия q, с которым мы уже рассматривали в предыдущей главе. Каждой из векторных норм (47) соответствует своя норма матрицы:

$$\|A\|_1 = \max_{1 \le j \le m} \sum_{i=1}^{m} |a_{ij}|, \quad \|A\|_2 \le \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{m} |a_{ij}|^2}, \quad \|A\|_\infty = \max_{1 \le i \le m} \sum_{j=1}^{m} |a_{ij}|. \tag{49}$$

Для вычисления нормы $\|A\|_1$ надо найти сумму модулей элементов каждого столбца матрицы A, а затем выбрать максимальную из этих сумм. Для вычисления нормы $\|A\|_\infty$ то же надо сделать не со столбцами, а со строками матрицы A. Заметим, что для нормы $\|A\|_2$ дана лишь оценка сверху, поскольку точное значение этой нормы вычисляется трудоемко.

Пример. Для матрицы

$$A = \begin{pmatrix} 0.1 & -0.4 & 0 \\ 0.2 & 0.3 & 0.1 \\ 0 & -0.1 & -0.3 \end{pmatrix}$$

вычислить $\|A\|_1$, $\|A\|_\infty$ и оценить $\|A\|_2$.

Решение. В соответствии с формулами (49)

$$\|A\|_1 = \max\{0.1 + 0.2 + 0;\ 0.4 + 0.3 + 0.1;\ 0 + 0.1 + 0.3\} = 0.8;$$

$$\|A\|_\infty = \max\{0.1 + 0.4 + 0;\ 0.2 + 0.3 + 0.1;\ 0 + 0.1 + 0.3\} = 0.6;$$

$$\|A\|_2 \le \sqrt{\sum_{i=1}^{3} \sum_{j=1}^{3} |a_{ij}|^2} =$$

$$= \sqrt{(0.1)^2 + (-0.4)^2 + 0^2 + (0.2)^2 + (0.3)^2 + (0.1)^2 + \mathbf{0}^2 + (-0.1)^2 + (-0.3)^2} =$$

$$= \sqrt{0.41} \approx 0.64.$$

Теперь можно сформулировать теорему, устанавливающую условие сходимости метода простой итерации и критерий обрыва итерационного процесса (45).

Теорема о сходимости итерационного процесса. Пусть выполнено условие

$$\|C\| < 1. \tag{50}$$

Тогда:

1) решение $\bar{x}$ системы (42) существует и единственно;

2) при произвольном векторе начального приближения $x^{(0)}$ метод простой итерации сходится к точному решению системы $\bar{x}$

$$\lim_{k\to\infty} x^{(k)} = \bar{x},$$

и справедлива следующая оценка абсолютной погрешности k-го приближения, обобщающая формулу (31),

$$\Delta = \left\|x^{(k)} - \bar{x}\right\| \le \frac{q}{1-q}\left\|x^{(k)} - x^{(k-1)}\right\|, \tag{51}$$

где в качестве коэффициента сжатия $q = \|C\|$ можно использовать любую норму матрицы (49), удовлетворяющую условию (50).

Несколько **замечаний** к теореме.

1. Из сходимости итераций по одной из норм следует и сходимость по другой норме, т. е., например, если $\left\|x^{(k)} - \bar{x}\right\|_1 \to 0$ при $k \to \infty$, то и $\left\|x^{(k)} - \bar{x}\right\|_2 \to 0$ при $k \to \infty$, и наоборот.

2. Приведем критерий достижения требуемой точности ε (критерий обрыва счета): вычисления можно прервать, как только абсолютная погрешность Δ, оцениваемая по формуле (51), станет меньше ε. Практически (в соответствии с нормой $\|x\|_\infty$ в формуле (47) следует проверить выполнение условия для каждой компоненты вектора k-го приближения

$$\frac{q}{1-q}\left\|x_i^{(k)} - x_i^{(k-1)}\right\| < \varepsilon, \quad i = 1, 2, \ldots, n. \tag{52}$$

3. Как и для любого сходящегося итерационного процесса, здесь выполняется условие самоисправления: вектор начального приближения $\boldsymbol{x}^{(0)}$ произволен, и при любом его выборе итерационный процесс сойдется к одному и тому же точному решению системы $\bar{\boldsymbol{x}}$. Обычно в качестве $\boldsymbol{x}^{(0)}$ берут вектор неоднородностей $\boldsymbol{f}$ в системе (44). Чем ближе начальное приближение к $\bar{\boldsymbol{x}}$, тем, при прочих равных условиях, потребуется меньше итераций для выполнения условий (53). Поэтому если каким-то образом стало известным грубо прикидочное решение системы (44), его выгодно использовать в качестве $\boldsymbol{x}^{(0)}$.

4. Простейший способ преобразования системы (43) к виду (44), удобному для итераций, состоит в том, чтобы в каждом уравнении (42) выразить соответствующую неизвестную через все остальные

$$\left.\begin{aligned} x_1 &= \frac{1}{a_{11}}\left(b_1 - a_{12}x_2 - a_{13}x_3 - \ldots - a_{1n}x_n\right), \\ x_2 &= \frac{1}{a_{22}}\left(b_2 - a_{21}x_1 - a_{23}x_3 - \ldots - a_{2\boldsymbol{n}}x_n\right), \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n &= \frac{1}{a_{nn}}\left(b_n - a_{n1}x_1 - a_{n2}x_2 - \ldots - a_{n\ n-1}x_{n-1}\right). \end{aligned}\right\}$$

Этот способ годится, если диагональные элементы матрицы A отличны от нуля:

$$a_{ii} \neq 0, \quad i = 1, 2, \ldots, n.$$

Реализацию метода простой итерации рассмотрим на примере.

Пример. Решить методом простой итерации систему

$$\left.\begin{aligned} 20.9x_1 + 1.2x_2 + 2.1x_3 + 0.9x_4 &= 21.70, \\ 1.2x_1 + 21.2x_2 + 1.5x_3 + 2.5x_4 &= 27.46, \\ 2.1x_1 + 1.5x_2 + 19.8x_3 + 1.3x_4 &= 28.76, \\ 0.9x_1 + 2.5x_2 + 1.3x_3 + 32.1\boldsymbol{x}_4 &= 49.72 \end{aligned}\right\} \qquad (53)$$

с точностью $\varepsilon = 10^{-3}$.

Решение. Приведем систему к виду, удобному для итераций:

$$\left.\begin{aligned} x_1 &= \frac{1}{20.9}\left(21.70 - 1.2x_2 - 2.1x_3 - 0.9x_4\right), \\ x_2 &= \frac{1}{21.2}\left(27.46 - 1.2x_1 - 1.5x_3 - 2.5x_4\right), \\ x_3 &= \frac{1}{19.8}\left(28.76 - 2.1x_1 - 1.5x_2 - 1.3x_4\right), \\ x_4 &= \frac{1}{32.1}\left(49.72 - 0.9x_1 - 2.5x_2 - 1.3x_3\right). \end{aligned}\right\} \qquad (54)$$

Матрица C (см. формулу (44), соответствующая системе (54) удовлетворяет ограничению на норму (50), если последнюю вычислять, например, как $\|C\|_\infty$ (49) — максимум построчной суммы модулей. В самом деле, вычислим эту сумму для каждой строки:

$$\sum_{j=1}^{4}\left|c_{1j}\right| = \left|c_{11}\right| + \left|c_{12}\right| + \left|c_{13}\right| + \left|c_{14}\right| = \frac{1.2 + 2.1 + 0.9}{20.9} \approx 0.20,$$

$$\sum_{j=1}^{4}\left|c_{2j}\right| \approx 0.24, \quad \sum_{j=1}^{4}\left|c_{3j}\right| \approx 0.25, \quad \sum_{j=1}^{4}\left|c_{4j}\right| \approx 0.15.$$

$$\|C\|_\infty = \max_{1\le i\le 4}\sum_{j=1}^{4}\left|c_{ij}\right| = \max\{0.20;\ 0.24;\ 0.25;\ 0.15\} = 0.25.$$

Итак, коэффициент сжатия $q = \|C\|_\infty = 0.25 < 1$. При таком значении q скорость сходимости итерационного процесса будет высокой, поскольку $q/(1-q) = 1/3$.

В качестве вектора начального приближения возьмем столбец неоднородностей в системе (54):

$$\boldsymbol{x}^{(0)} = \begin{pmatrix} 21.70/20.9 \\ 27.46/21.2 \\ 28.76/19.8 \\ 49.72/32.1 \end{pmatrix} \approx \begin{pmatrix} 1.04 \\ 1.30 \\ 1.45 \\ 1.55 \end{pmatrix}.$$

Итерации будем продолжать до тех пор, пока оценка погрешности

$$\frac{q}{1-q}\left\|\boldsymbol{x}_i^{(k)} - \boldsymbol{x}_i^{(k-1)}\right\| = \tfrac{1}{3}\left\|\boldsymbol{x}_i^{(k)} - \boldsymbol{x}_i^{(k-1)}\right\|$$

не станет меньше $\varepsilon = 10^{-3}$ для всех компонент вектора.

Первая итерация (ср. с системой (54):

$$x_1^{(1)} = \frac{1}{20.9}\left(21.70 - 1.2x_2^{(0)} - 2.1x_3^{(0)} - 0.9x_4^{(0)}\right) =$$

$$= \frac{1}{20.9}\cdot(21.70 - 1.2\cdot 1.30 - 2.1\cdot 1.45 - 0.9\cdot 1.55) = 0.75,$$

$$x_2^{(1)} = \frac{1}{21.2}\left(27.46 - 1.2x_1^{(0)} - 1.5x_3^{(0)} - 2.5x_4^{(0)}\right) =$$

$$= \frac{1}{21.2}\cdot(27.46 - 1.2\cdot 1.04 - 1.5\cdot 1.45 - 2.5\cdot 1.55) = 0.95,$$

$$x_3^{(1)} = \frac{1}{19.8}\left(28.76 - 2.1x_1^{(0)} - 1.5x_2^{(0)} - 1.3x_4^{(0)}\right) =$$

$$= \frac{1}{19.8}\cdot(28.76 - 2.1\cdot 1.04 - 1.5\cdot 1.30 - 1.3\cdot 1.55) = 1.14,$$

$$x_4^{(1)} = \frac{1}{32.1}\left(49.72 - 0.9x_1^{(0)} - 2.5x_2^{(0)} - 1.3x_3^{(0)}\right) =$$

$$= \frac{1}{32.1}\cdot(49.72 - 0.9\cdot 1.04 - 2.5\cdot 1.30 - 1.3\cdot 1.45) = 1.36$$

итак, $\boldsymbol{x}^{(1)} = \begin{pmatrix} 0.75 \\ 0.95 \\ 1.14 \\ 1.36 \end{pmatrix}$.

(Вычисления пока можно проводить с небольшим числом знаков после запятой, поскольку трудно ожидать высокой точности от вектора первого приближения.)

Вторая итерация $\boldsymbol{x}^{(2)} = \begin{pmatrix} 0.8106 \\ 1.0118 \\ 1.2117 \\ 1.4077 \end{pmatrix}$.

Третья итерация $\boldsymbol{x}^{(3)} = \begin{pmatrix} 0.7978 \\ 0.9977 \\ 1.1975 \\ 1.3983 \end{pmatrix}$.

Четвертая итерация $\boldsymbol{x}^{(4)} = \begin{pmatrix} 0.8004 \\ 1.0005 \\ 1.2005 \\ 1.4003 \end{pmatrix}$.

Оценим точность системы (54), достигнутую после четвертой итерации:

$$\frac{1}{3}\left\|\boldsymbol{x}_1^{(4)} - \boldsymbol{x}_1^{(3)}\right\| = 0.0009, \quad \frac{1}{3}\left\|\boldsymbol{x}_2^{(4)} - \boldsymbol{x}_2^{(3)}\right\| = 0.0009,$$

$$\frac{1}{3}\left\|\boldsymbol{x}_3^{(4)} - \boldsymbol{x}_3^{(3)}\right\| = 0.001, \quad \frac{1}{3}\left\|\boldsymbol{x}_4^{(4)} - \boldsymbol{x}_4^{(3)}\right\| = 0.0007.$$

Поскольку максимальная из оценок (третья) не меньше $\varepsilon = 0.001$, продолжаем вычисления.

Пятая итерация $\boldsymbol{x}^{(5)} = \begin{pmatrix} 0.7999 \\ 0.9999 \\ 1.1999 \\ 1.3999 \end{pmatrix}$.

Оценим точность после пятой итерации:

$$\frac{1}{3}\left\|\boldsymbol{x}_1^{(5)} - \boldsymbol{x}_1^{(4)}\right\| = 1{,}7 \cdot 10^{-4}, \quad \frac{1}{3}\left\|\boldsymbol{x}_2^{(5)} - \boldsymbol{x}_2^{(4)}\right\| = 2 \cdot 10^{-4},$$

$$\frac{1}{3}\left\|\boldsymbol{x}_3^{(5)} - \boldsymbol{x}_3^{(4)}\right\| = 2 \cdot 10^{-4}, \quad \frac{1}{3}\left\|\boldsymbol{x}_4^{(5)} - \boldsymbol{x}_4^{(4)}\right\| = 1{,}3 \cdot 10^{-4}.$$

Абсолютная погрешность

$$\Delta_5 = \left\|\boldsymbol{x}^{(5)} - \bar{\boldsymbol{x}}\right\|_\infty \le \frac{\frac{1}{4}}{1 - \frac{1}{4}}\left\|\boldsymbol{x}^{(5)} - \boldsymbol{x}^{(4)}\right\|_\infty = \frac{1}{3}\max_{1 \le i \le m}\left|x_i^{(5)} - x_i^{(4)}\right| = 2 \cdot 10^{-4} < 10^{-3} = \varepsilon.$$

Итак, абсолютная погрешность впервые стала меньше требуемой по условию задачи. Вычисления заканчиваем и записываем ответ, округляя $\boldsymbol{x}^{(5)}$ до трех верных десятичных знаков, которые гарантированы достигнутой точностью $\varepsilon = 0.001$:

$$\bar{\boldsymbol{x}} \approx \boldsymbol{x}^{(5)} = \begin{pmatrix} 0.800 \\ 1.000 \\ 1.200 \\ 1.400 \end{pmatrix}.$$

Для информации отметим, что точное решение системы (53) есть $x_1 = 0.8$, $x_2 = 1.0$, $x_3 = 1.2$, $x_4 = 1.4$.

Замечания. 1. При практической реализации метода простой итерации, например в MathCad, рекомендуется ввести четыре функции нескольких переменных, соответствующие правым частям уравнений (54), и на каждой итерации вызывать эти функции по их именам, не переписывая в явном виде саму систему (54).

2. Понятно, что не всегда простейший переход от системы (53) к (54) обеспечивает выполнение условия сходимости (50) по какой-нибудь норме. Могут потребоваться дополнительные элементарные преобразования строк в исходной системе (53). Пример таких преобразований приведен ниже, при рассмотрении метода Зейделя.

Методы интерполирования и экстраполяции функций

Приближением (аппроксимацией) функции $f(x)$ называется отыскание функции $g(x)$, близкой в некотором смысле к $f(x)$. Аппроксимирующая функция $g(x)$ должная быть «проще» исходной. Как понимается близость функций и в чем критерий простоты, об этом речь пойдет ниже.

Аппроксимация может потребоваться в следующих случаях:

1) известны, например из эксперимента, значения функции $f(x_1)=y_1$, $f(x_2)=y_2$,.., $f(x_n)=y_n$ (итак, функция $y=f(x)$ задана таблично). Требуется найти значение $f(x)$ при таком значении аргумента x^*, которого нет среди узлов $x_1, x_2, \ldots x_n$, но сделать это по каким-либо причинам затруднительно [1]. В таком случае можно найти аппроксимирующую функцию $g(x)$; если она «близка» к $f(x)$ на множестве узлов $X=\{x_1, x_2, \ldots, x_n\}$, то и в нужной точке x^*, вероятно, $f(x^*)\approx g(x^*)$;

2) функция $f(x)$ задана аналитически, т.е. формулой, но эта формула слишком сложна [2] для регулярного использования. И в этом случае выгодно аппроксимировать $f(x)$ более простой функцией $g(x)$ и все расчеты выполнять с ней.

---

[1] Например, экспериментальная установка, на которой выполнены измерения, уже разобрана.

[2] Известно, например, что интеграл вида $\int \frac{P_n(x)}{Q_m(x)}\,dx$ от любой дробно-рациональной функции ($P_n(x)$ и $Q_m(x)$ — полиномы) всегда берущийся, т.е. первообразная выражается в конечном виде через элементарные функции, но формула для первообразной может быть очень громоздкой: $\int \frac{dx}{x^3\left(x^3+a^3\right)^2} = -\frac{1}{3a^3x^2\left(x^3+a^3\right)} - \frac{5}{6a^6x^2} - \frac{5}{18a^8}\ln\frac{(x+a)^2}{x^2-ax+a^2} - \frac{5}{3a^8\sqrt{3}}\operatorname{arctg}\frac{2x-a}{a\sqrt{3}}$. Это отнюдь не предел сложности!

Какие функции наиболее «просты» и в силу этого удобны в качестве аппроксимирующих? Чаще всего используются полиномы (многочлены)

$$P_n(x) = a_0 + a_1 x + \ldots + a_n x^n . \qquad (2)$$

Действительно, полиномы легко складывать, умножать и делить; их можно элементарно дифференцировать и интегрировать.

Иногда (но не в данном курсе) применяют обобщенные полиномы вида

$$a_0\varphi_0(x) + a_1\varphi_1(x) + \ldots + a_n\varphi_n(x),$$

где функции $\varphi_i(x)$, $i = 1, 2, \ldots, n$, предполагаются линейно независимыми. Например, функции $\{1, \sin x, \cos x, \sin 2x, \cos 2x, \ldots, \sin nx, \cos nx\}$ тоже обладают удобными свойствами и используются для разложения произвольных $f(x)$ при весьма общих условиях в тригонометрические ряды Фурье.

Рассмотрим теперь некоторые подходы к понятию близости функций.

1. Интерполяция. Требуется найти полином $P_n(x)$, принимающий те же значения, что и аппроксимируемая функция $f(x)$, в $(n+1)$ узле

$$P_n(x_i) = f(x_i),\ i = 0, 1, \ldots, n.$$

В таком случае полином $P_n(x)$ называется интерполяционным, а точки $x_0, x_1, \ldots, x_n$ — узлами интерполяции.

Если число узлов велико, то отыскание интерполяционного полинома, как мы увидим далее, будет трудоемким. Кроме того, точное равенство $P_n(x_i) = f(x_i)$ может оказаться бессмысленным требованием, если сами значения $f(x_i)$ аппроксимируемой функции в узлах получены из эксперимента и потому заведомо неточно.

2. Наилучшее приближение. Пусть функция $f(x)$ задана таблично в узлах $x_0, x_1, \ldots, x_n$. Подберем полином $P_n(x) = a_0 + a_1 x + \ldots + a_n x^n$ так, чтобы сумма квадратов разностей значения аппроксимируемой функции и полинома по всем узлам минимизировалась

$$\sum_{i=0}^{n} \left(f(x_i) - P_n(x_i)\right)^2 \to \min .$$

Из данного условия определяются коэффициенты $a_0, a_1, \ldots, a_n$ искомого полинома.

На этом пути мы пришли бы к знаменитому методу наименьших квадратов, но рассматривать его здесь мы не будем.

Займемся интерполяцией. Итак, пусть функция $f(x)$ задана таблично в $(n+1)$ узле

$$f(x_0) = y_0,\ f(x_1) = y_1,\ \ldots,\ f(x_n) = y_n.$$

Требуется найти интерполяционный полином, такой, что

$$P_n(x_i) = f(x_i),\ i = 0, 1, \ldots, n. \tag{3}$$

Геометрически это означает, что график полинома проходит через точки $M_i(x_i, y_i),\ i = 0,1,\ldots,n$ (рис. 1).

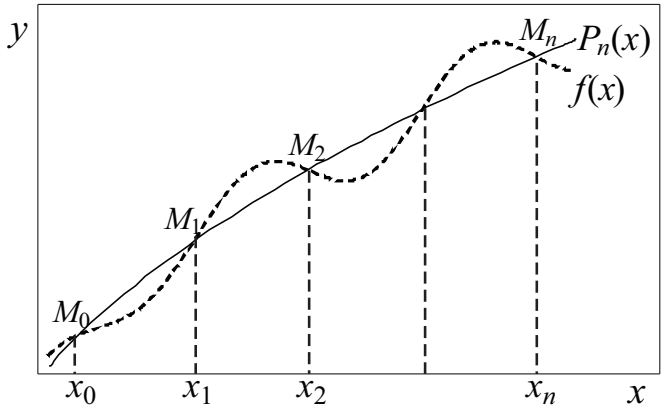

Рис. 1. Графики аппроксимируемой функции и ее интерполяционного полинома

Интерполяционный полином используется для приближенного вычисления значений функции $f(x)$ в точке, отличной от узлов интерполяции, $f(x) \approx P_n(x)$. Если значение x лежит между узлами интерполяции, то приближенное отыскание значения $f(x)$ называется **интерполированием** (в узком смысле); если же значение x лежит левее или правее всех узлов, то говорят об интерполировании в широком смысле (или **экстраполировании**).

Степень интерполяционного полинома, построенного по $(n+1)$ точке, в исключительных случаях может оказаться меньше n. Например, если все точки $M_i(x_i, y_i)$ лежат на одной прямой (наклонной или го-

ризонтальной), то полином будет иметь первую (соответственно нулевую) степень. Понятно, что такие случаи крайне редки.

В общем случае по заданным значениям $f(x)$ в $(n+1)$ узле конструируется полином степени n, притом единственным образом. Покажем это. Подставляя в общий вид полинома (2) условия (3), получим систему из $(n+1)$ линейного уравнения

$$\begin{cases} a_0 + a_1 x_0 + \ldots + a_n x_0^n = y_0, \\ a_0 + a_1 x_1 + \ldots + a_n x_1^n = y_1, \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ a_0 + a_1 x_n + \ldots + a_n x_n^n = y_n \end{cases}$$

с $(n+1)$ неизвестными — коэффициентами полинома $a_0, a_1, \ldots, a_n$. Решать эту систему можно, например, по правилу Крамера. Если главный определитель системы (составленный из коэффициентов при неизвестных) отличен от нуля, то система имеет решение и притом единственное. В нашем случае главный определитель имеет специальный вид (так называемый определитель Вандермонда):

$$\Delta = \begin{vmatrix} 1 & x_0 & \ldots & x_0^n \\ 1 & x_1 & \ldots & x_1^n \\ \cdot & \cdot & \cdot & \cdot \\ 1 & x_n & \ldots & x_n^n \end{vmatrix}.$$

Доказывается, что $\Delta \neq 0$, поскольку узлы интерполяции $x_0, x_1, \ldots, x_n$ — различные числа. Итак, решение системы — коэффициенты интерполяционного полинома $a_0, a_1, \ldots, a_n$ — существует и единственно. Поэтому по данным значениям $f(x)$ в $(n+1)$ узле можно построить полином $P_n(x)$ степени n, притом единственным образом. Есть несколько способов построения, приводящих к одинаковому результату. Мы рассмотрим простейший способ, предложенный Лагранжем (1736–1813).

Интерполяционный полином Лагранжа

Для системы узлов $x_0, x_1, \ldots, x_n$ введем коэффициенты Лагранжа вида

$$L_n^i(x) = \frac{(x-x_0)(x-x_1)\cdot\ldots\cdot(x-x_{i-1})(x-x_{i+1}\cdot), \quad \cdot(x-x_n)}{(x_i-x_0)(x_i-x_1)\cdot\ldots\cdot(x_i-x_{i-1})(x_i-x_{i+1}\cdot), \quad \cdot(x_i-x_n)}.$$

Здесь индекс i может принимать значения $0,1,\ldots,n$. В числителе каждый сомножитель представляет собой разность переменного x и значения одного из узлов (за исключением i-го узла, что отмечено верхним индексом в обозначении функции L_n^i). Знаменатель формально отличается от числителя тем, что вместо переменного x подставлено значение пропущенного, i-го, узла интерполяции x_i. После упрощений становится понятно, что коэффициент Лагранжа $L_n^i(x)$ является полиномом n-й степени (что отмечено нижним индексом в обозначении L_n^i).

При подстановке значения j-го узла в качестве аргумента коэффициента Лагранжа получится

$$L_n^i(x_j)=\begin{cases}1, & j=i,\\ 0, & j\neq i.\end{cases}$$

В таком случае полином Лагранжа

$$L_n(x)=\sum_{i=0}^{n} f(x_i)L_n^i(x)=f(x_0)L_n^0(x)+f(x_1)L_n^1(x)+\ldots+f(x_n)L_n^n(x) \quad (4)$$

удовлетворяет условиям (3)

$$L_n(x_i)=f(x_i),\ i=0,1,\ldots,n,$$

и имеет степень n, т. е. является искомым интерполяционным полиномом.

Пример. Функция $y=f(x)$ задана таблично своими значениями в четырех узлах:

i	0	1	2	3
узлы x_i	-1	0	2	5
$y_i=f(x_i)$	1	-3	2	4

Построить для $y=f(x)$ интерполяционный полином Лагранжа и, пользуясь им, приближенно найти значение y в точке $x=1$, которой нет среди узлов.

Решение. Применяя формулу (4), получим

$$L_3(x) = y_0 \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} + y_1 \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} +$$

$$+ y_2 \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} + y_3 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} =$$

$$= 1\frac{(x-0)(x-2)(x-5)}{(-1-0)(-1-2)(-1-5)} - 3\frac{(x+1)(x-2)(x-5)}{(0+1)(0-2)(0-5)} +$$

$$+2\frac{(x+1)(x-0)(x-5)}{(2+1)(2-0)(2-5)} + 4\frac{(x+1)(x-0)(x-2)}{(5+1)(5-0)(5-2)} = -\frac{19}{45}x^3 + \frac{233}{90}x^2 - \frac{89}{90}x - 3.$$

Тогда

$$f(1) \approx L_3(1) = -\frac{19}{45} + \frac{233}{90} - \frac{89}{90} - 3 = -82/45 = -1.82.$$

Итак, ответ получен, но вопрос о его точности пока остается открытым. Заметим, что имела место интерполяция в узком смысле, поскольку точка $x = 1$ лежит между узлами интерполяции.

Погрешность интерполяционного полинома Лагранжа

Погрешностью интерполяции называется модуль разности значений аппроксимируемой функции и ее интерполяционного полинома

$$R_n(x) = \left| f(x) - L_n(x) \right|. \qquad (5)$$

Применимость этой формулы ограничена. Действительно, если значение $f(x)$ известно точно, то необходимость в аппроксимации отпадает и вопрос о погрешности интерполяции беспредметен.

По определению формула (5) дает точное значение погрешности интерполяции. Для практики удобнее оказывается приближенная формула, оценивающая погрешность сверху. Пусть узлы интерполяции $x_0, x_1, \ldots, x_n$ все принадлежат отрезку $[a, b]$. Предположим, что аппроксимируемая функция имеет производную $(n+1)$-го порядка на этом отрезке. Без доказательства

$$R_n(x)=\frac{\left|f^{(n+1)}(\xi)\right|}{(n+1)!}\left|(x-x_0)(x-x_1)\cdot\ldots\cdot(x-x_n)\right|, \qquad (6)$$

где ξ — некоторая (вообще говоря неизвестная) точка, лежащая между узлами интерполяции, $\xi\in[a,b]$; в ней вычисляется $(n+1)$-я производная аппроксимируемой функции. Вычисление промежуточной точки ξ и производной (высокого порядка!) настолько трудны, что делают (точную) формулу (6) фактически неприменимой. Ее можно упростить: если каким-либо способом оценить сверху $(n+1)$-ю производную

$$\max_{x\in[a,b]}\left|f^{(n+1)}(x)\right|=M_{n+1}$$

$(M_{n+1}-\text{число})$, то получим следующую оценку погрешности интерполяции

$$R_n(x)\le\frac{M_{n+1}}{(n+1)!}\left|(x-x_0)(x-x_1)\cdot\ldots\cdot(x-x_n)\right|. \qquad (7)$$

Пример. Функция $f(x)=\ln x$ задана таблично своими значениями в узлах $x_0=100$, $x_1=101$, $x_2=102$, $x_3=103$, и по табличным данным построен интерполяционный полином Лагранжа $L_3(x)$, который применен для приближенного вычисления $\ln 100{,}5$. Оценить, с какой точностью получается это значение.

Решение. Требуется оценить погрешность интерполяции $R_3(100.5)$. Согласно формуле (7)

$$R_3(x)\le\frac{M_4}{4!}\left|(x-x_0)(x-x_1)(x-x_2)(x-x_3)\right|,$$

где $M_4=\max\limits_{x\in[a,b]}\left|f^{(n+1)}(x)\right|=\max\limits_{x\in[100,103]}\left|\ln^{(4)}(x)\right|$ (минимальный отрезок, содержащий все четыре узла, есть отрезок $[100,103]$). Четвертая производная логарифма $\ln^{(4)}(x)=-6/x^4$, поэтому $M_4=\max\limits_{x\in[100,103]}\left|-\frac{6}{x^4}\right|=\frac{6}{100^4}$.

Оценка погрешности интерполяции

$$R_3(100.5)=\left|\ln 100.5-L_3(100.5)\right|\le$$

$$\le\frac{6}{100^4 4!}\left|(100.5-100)\cdot(100.5-101)\cdot(100.5-102)\cdot(100.5-103)\right|<10^{-8}.$$

Итак, вычисление $\ln 100.5$ с помощью интерполяционного полинома Лагранжа $L_3(x)$ даст ответ с восемью верными десятичными знаками, т. е. очень точный ответ.

Численное дифференцирование и интегрирование функций

Численное дифференцирование

Напомним сначала, как выполняется дифференцирование, т. е. нахождение производной, в математическом анализе, и выясним, почему такой подход не всегда возможен.

Пусть имеется функция, заданная аналитически, т. е. формулой $y = f(x)$. Начальное значение аргумента есть x_0; при этом функция принимает значение $y_0 = f(x_0)$. Если аргумент испытает приращение Δx так, что новое значение аргумента $x_1 = x_0 + \Delta x$, то ему будет соответствовать новое значение функции $y_1 = f(x_1)$. Приращение функции в точке x_0 есть $\Delta f(x_0) = f(x_1) - f(x_0)$. Отношение $\Delta f(x_0)/\Delta x$ приращения функции к приращению аргумента показывает среднюю скорость изменения функции на отрезке от x_0 до x_1. Если теперь устремить приращение аргумента к нулю, то предел отношения (если этот предел существует) по определению равен значению производной функции в точке x_0

$$f'(x_0) = \lim_{\Delta x \to 0} \frac{\Delta f(x_0)}{\Delta x}. \qquad (8)$$

К сожалению, эта формула применима не всегда. Если функция $f(x)$ задана табличнo своими значениями в узлах $x_0, x_1, ..$, то приращение аргумента $\Delta x = x_i - x_{i-1}$, как бы мало оно ни было, всегда конечно, и его нельзя устремить к нулю.

x	x_0	x_1	x_2	x_3	...
y	y_0	y_1	y_2	y_3	...

Например, индекс РТС, являющийся основным показателем фондового рынка России, рассчитывается каждые 15 с. Можно вычислить среднюю скорость изменения индекса на каждом временно́м интервале, но найти точное значение производной индекса РТС по времени невозможно.

Итак, к ч и с л е н н о м у дифференцированию приходится прибегать в тех случаях, когда функция $f(x)$, которую нужно продифференцировать, задана таблично. Кроме того, даже если функция задана аналитически, но очень сложной формулой[3], нахождение производной по определению (8) может оказаться проблематичным — вычислить значения $f(x_i)$ затруднительно.

Идея численного дифференцирования несложна: вместо функции $f(x)$ рассматривается некоторая «близкая» к ней функция[4] $g(x)$, определяемая достаточно простой формулой, и поскольку $f(x) \approx g(x)$, постольку их производные, вероятно, близки

$$f'(x) \approx g'(x);$$

это должно быть верно и для вторых производных

$$f''(x) \approx g''(x),$$

и т. д.

В действительности эта идея очень рискованна: близость функций отнюдь не предопределяет близость их производных. В качестве примера рассмотрим две функции: $f(x)$ и $g(x) = f(x) + \frac{1}{n}\sin n^2 x$. Поскольку $\left|f(x) - g(x)\right| = \left|\frac{1}{n}\sin n^2 x\right| \xrightarrow{n\to\infty} 0$, разность функций при достаточно больших n сколь угодно мала, т. е. ф у н к ц и и б л и з к и. Однако разность их производных $\left|f'(x) - g'(x)\right| = \left|f'(x) - \left[f'(x) + n\cos n^2 x\right]\right| = \left|n\cos n^2 x\right|$ не мала, т. к. $\max\limits_{x\in(-\infty,\,\infty)} \left|n\cos n^2 x\right| = n$, т. е. производные н е б л и з к и.

Это можно показать проще, опираясь на геометрический смысл производной $f'(x_0)$: она равна тангенсу угла наклона касательной к графику функции $y = f(x)$ в точке с абсциссой x_0 (рис. 2).

Функции $f(x)$ и $g(x)$ близки, т. к. ординаты точек двух графиков при одинаковых абсциссах мало отличаются. Но направления касательных, проведенных к графикам (см., например, точки A и B), резко различны, т. е. производные $f'(x)$ и $g'(x)$ н е б л и з к и.

---

[3] См. сноску 2.

[4] Как мы помним, говорят об аппроксимации $f(x)$ с помощью $g(x)$.

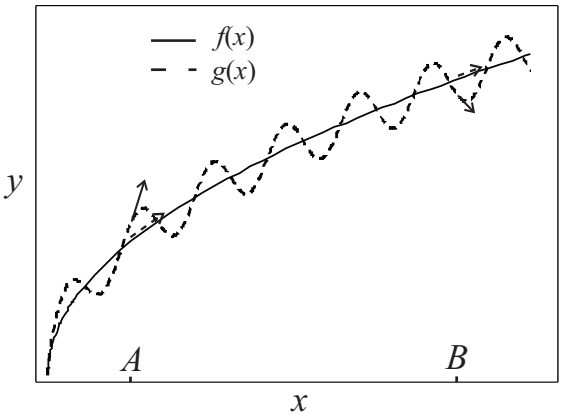

Рис. 2. Сопоставление двух «близких» функций[5]

Итак, вообще говоря, «наивная» идея (2)–(3) неверна, но если в качестве аппроксимирующей функции $g(x)$ используется интерполяционный полином Лагранжа, построенный по значениям $f(x)$ во многих узлах с небольшим расстоянием между ними, то приближенно вычислить $f'(x)$, заменяя ее на $g'(x)$, все же возможно, и возникающая при этом погрешность численного дифференцирования может быть оценена.

Пусть функция $f(x)$ задана таблично своими значениями в равноотстоящих узлах $x_0, x_1, x_2, \ldots, x_n$. По этим значениям построим полином Лагранжа $L_n(x)$, и тогда $f'(x) \approx L_n'(x)$, $f''(x) \approx L_n''(x)$ и т. д. Приведем вычисления для простейшего случая $n = 2$.

Формулы численного дифференцирования для трех равноотстоящих узлов

Имеются три узла x_0, x_1, x_2, удаленные друг от друга на расстояние h (шаг) (рис. 3). Интерполяционный полином Лагранжа

$$L_2(x) = y_0 \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} + y_1 \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} + y_2 \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}.$$

[5] Функции $f(x)$ и $g(x)$ близки, т. к. ординаты точек двух графиков при одинаковых абсциссах мало отличаются. Но направления касательных, проведенных к графикам (см., например, точки A и B), резко различны, т. е. производные $f'(x)$ и $g'(x)$ не близки.

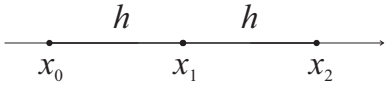

Рис. 3. Три равноотстоящих узла

Введем вспомогательную переменную $q = \dfrac{x - x_0}{h}$, тогда $x = x_0 + hq$, и получаем

$$L_2(x) = \frac{1}{2} y_0 (q-1)(q-2) - y_1 q(q-2) + \frac{1}{2} y_2 q(q-1).$$

Полином Лагранжа выглядит теперь компактнее, но его нельзя непосредственно продифференцировать по x, поскольку аргументом стала q. По правилу дифференцирования сложной функции $(L_2)_x' = (L_2)_q' q_x' = \dfrac{1}{h}(L_2)_q'$, получаем следующее приближенное соотношение для производной:

$$f'(x) \approx L_2'(x) = \frac{1}{h}\left[\frac{1}{2} y_0 (2q-3) - y_1 (2q-2) + \frac{1}{2} y_2 (2q-1)\right]. \qquad (9)$$

Аналогично — для второй производной

$$f''(x) \approx L_2''(x) = \frac{1}{h^2}\left[y_0 - 2y_1 + y_2\right]. \qquad (10)$$

Подстановка $x = x_0 + hq$ при $q = 0, 1, 2$ дает соответственно $x = x_0$, x_1, x_2. Поэтому из выражений (9) и (10) можно при таких значениях q найти приближенные выражения для производных в каждом из трех узлов (последнее слагаемое в каждой из нижеследующих формул — погрешность численного дифференцирования, которую мы приводим без доказательства):

первая производная

$$f'(x_0) = \frac{1}{2h}(-3y_0 + 4y_1 - y_2) + \frac{h^2}{3} f'''(\xi), \qquad (11)$$

$$f'(x_1) = \frac{1}{2h}(-y_0 + y_2) - \frac{h^2}{6} f'''(\xi), \qquad (12)$$

$$f'(x_2) = \frac{1}{2h}(y_0 - 4y_1 + 3y_2) + \frac{h^2}{3} f'''(\xi); \qquad (13)$$

вторая производная

$$f''(x_0)=\frac{1}{h^2}(y_0-2y_1+y_2)-hf'''(\xi),$$

$$f''(x_1)=\frac{1}{h^2}(y_0-2y_1+y_2)-\frac{h^2}{12}f^{\mathrm{IV}}(\xi),$$

$$f''(x_2)=\frac{1}{h^2}(y_0-2y_1+y_2)+hf'''(\xi)$$

(поскольку в формулу (10) переменная q не вошла, вторая производная в разных узлах отличается только погрешностью). Во всех формулах ξ есть некоторая (неизвестная) промежуточная точка, лежащая между узлами.

Мы подробно рассмотрели способ получения формул численного дифференцирования для случая трех равноотстоящих узлов. Приведем без вывода аналогичные формулы для случая четырех равноотстоящих узлов x_0, x_1, x_2, x_3, находящихся на расстоянии h друг от друга.

Формулы численного дифференцирования для четырех равноотстоящих узлов

Первая производная

$$f'(x_0)=\frac{1}{6h}(-11y_0+18y_1-9y_2+2y_3)-\frac{h^3}{4}f^{\mathrm{IV}}(\xi), \tag{14}$$

$$f'(x_1)=\frac{1}{6h}(-2y_0-3y_1+6y_2-y_3)+\frac{h^3}{12}f^{\mathrm{IV}}(\xi), \tag{15}$$

$$f'(x_2)=\frac{1}{6h}(y_0-6y_1+3y_2+2y_3)-\frac{h^3}{12}f^{\mathrm{IV}}(\xi), \tag{16}$$

$$f'(x_3)=\frac{1}{6h}(-2y_0+9y_1-18y_2+11y_3)+\frac{h^3}{4}f^{\mathrm{IV}}(\xi). \tag{17}$$

Вторая производная

$$f''(x_0)=\frac{1}{h^2}(2y_0-5y_1+4y_2-y_3)+\frac{11}{12}h^2f^{\mathrm{IV}}(\xi), \tag{18}$$

$$f''(x_1)=\frac{1}{h^2}(y_0-2y_1+y_2)-\frac{1}{12}h^2f^{\mathrm{IV}}(\xi), \tag{19}$$

$$f''(x_2) = \frac{1}{h^2}(y_1 - 2y_2 + y_3) - \frac{1}{12}h^2 f^{IV}(\xi), \tag{20}$$

$$f''(x_3) = \frac{1}{h^2}(-y_0 + 4y_1 - 5y_2 + 2y_3) + \frac{11}{12}h^2 f^{IV}(\xi). \tag{21}$$

Замечания к формулам численного дифференцирования:

1) оценка погрешности уже первой производной требует предварительного знания третьей (формулы (11)–(13) и даже четвертой (!) (формулы (14)–(17) производной и притом в заранее неизвестной точке ξ. Тем самым погрешность численного дифференцирования оценивается настолько сложно, что в лабораторной работе эта оценка не предполагается;

2) выполняется некоторое правило симметрии, позволяющее проверить правильность написания формул, а именно: коэффициенты при y_i для первой производной в узлах, симметричных относительно центрального, повторяются с обратным знаком, если прочитать формулу справа налево — ср. формулы (11) и (13), (14) и (17), (15) и (16). Для второй производной — то же, но без смены знака (ср. формулы (18) и (21), (19) и (20);

3) в узлах, близких к центральному, формулы численного дифференцирования проще и отличаются большей точностью (меньшей погрешностью), чем в крайних узлах;

4) с увеличением количества узлов точность формул численного дифференцирования повышается (это неудивительно, ведь возрастает количество информации, которая имеется у нас о дифференцируемой функции);

5) с ростом порядка производной точность формул уменьшается (погрешность растет). Это связано с тем, что дифференцирование ухудшает свойства функции (дифференцируемая всюду функция может иметь производную функцию, которая не везде будет дифференцируемой; непрерывная функция может перейти в разрывную).

Численное интегрирование

Если функция $f(x)$ задана аналитически (формулой) и ее первообразная $F(x)$ является элементарной функцией, то определенный интеграл $\int_a^b f(x)\mathrm{d}x$ вычисляется по формуле Ньютона – Лейбница

$$\int_a^b f(x)\mathrm{d}x = F(x)\Big|_a^b = F(b) - F(a).$$

Существуют ситуации, когда этой формулой невозможно или затруднительно воспользоваться:

1) подынтегральная функция $f(x)$ задана графически или таблично; тогда первообразная $F(x)$ не существует;

2) подынтегральная функция $f(x)$ задана аналитически, но интеграл $\int f(x)\mathrm{d}x$ неберущийся, т.е. не выражается в конечном виде через элементарные функции (известно, что многие важные интегралы, часто встречающиеся в практических приложениях, таковы, в качестве примера приведем $\int e^{-x^2}\mathrm{d}x$, $\int \frac{\sin x}{x}\mathrm{d}x$);

3) подынтегральная функция $f(x)$ задана аналитически и интеграл $\int f(x)\mathrm{d}x$ берущийся, но первообразная $F(x)$ слишком громоздка (см. примечание в начале гл. 3 о численном дифференцировании).

Во всех этих случаях приходится прибегать к приближенному, численному нахождению определенного интеграла. Для этого подынтегральную функцию $f(x)$ заменяют другой, «близкой» к ней функцией, которая легко интегрируется.

Формула Ньютона — Котеса

В качестве функции, «близкой» к $f(x)$, возьмем интерполяционный полином Лагранжа $L_m(x)$, совпадающий с $f(x)$ в узлах интерполяции $x_0, x_1, \ldots, x_m$, лежащих на отрезке интегрирования $[a, b]$. Полином Лагранжа имеет вид

$$L_m(x) = \sum_{i=0}^{m} f(x_i)\, L_m^i(x), \quad m = 0, 1, 2, \ldots,$$

где $L_m^i(x)$ — коэффициенты Лагранжа (полиномы степени m)

$$L_0^0(x) \equiv 1,$$

$$L_m^i(x) = \frac{(x - x_0)(x - x_1)\ldots(x - x_{i-1})(x - x_{i+1})\quad (x - x_m)}{(x_i - x_0)(x_i - x_1)\ldots(x_i - x_{i-1})(x_i - x_{i+1})\ldots(x_i - x_m)}.$$

Если полином Лагранжа «близок» к функции $f(x)$, то и интегралы от них тоже должны быть близки[6]:

$$\int_a^b f(x)\,\mathrm{d}x \approx \int_a^b L_m(x)\,\mathrm{d}x = \sum_{i=0}^{m} f(x_i) \int_a^b L_m^i(x)\,\mathrm{d}x.$$

Вводя так называемые коэффициенты Котеса (1682–1716)

$$c_m^i = \int_a^b L_m^i(x)\,\mathrm{d}x,$$

получаем формулу Ньютона – Котеса порядка m

$$\int_a^b f(x)\,\mathrm{d}x \approx \int_a^b L_m(x)\,\mathrm{d}x = \sum_{i=0}^{m} c_m^i\, f(x_i).$$

Она позволяет приближенно представить значение определенного интеграла в виде линейной комбинации значений подынтегральной функции в узлах интерполяции.

---

[6] В случае численного дифференцирования, как мы знаем, такое рассуждение оказалось рискованным, но здесь оно вполне оправданно. Причина в том, что дифференцирование *ухудшает* свойства функции (производная дифференцируемой всюду функции может стать недифференцируемой в данной точке, производная непрерывной функции — разрывной функцией), а интегрирование, как обратная к дифференцированию операция, *улучшает* свойства функции.

Пример. Вычислить коэффициенты Котеса c_1^0 и c_1^1.

Решение. Пусть значения функции $f(x)$ заданы в двух узлах: $x_0 = a$ и $x_1 = b$. В таком случае функцию можно аппроксимировать полиномом Лагранжа первой степени

$$f(x) \approx L_1(x) = f(x_0) L_1^0(x) + f(x_1) L_1^1(x) =$$

$$= f(x_0) \frac{x - x_1}{x_0 - x_1} + f(x_1) \frac{x - x_0}{x_1 - x_0} = f(a) \frac{x - b}{a - b} + f(b) \frac{x - a}{b - a}.$$

Интеграл от аппроксимируемой функции

$$\int_{x_0}^{x_1} f(x) \mathrm{d}x \approx \frac{f(a)}{a - b} \int_a^b (x - b) \mathrm{d}x + \frac{f(b)}{b - a} \int_a^b (x - a) \mathrm{d}x =$$

$$= f(a) \frac{b - a}{2} + f(b) \frac{b - a}{2} = \frac{b - a}{2} \left[f(a) + f(b) \right].$$

Отсюда получаем коэффициенты Котеса — весовые коэффициенты при значениях аппроксимируемой функции $f(a)$ и $f(b)$

$$c_1^0 = c_1^1 = \frac{b - a}{2}.$$

Отметим очевидный смысл последней формулы: определенный интеграл (геометрически равный площади криволинейной трапеции) в самом грубом приближении подсчитывается как площадь трапеции (рис. 4).

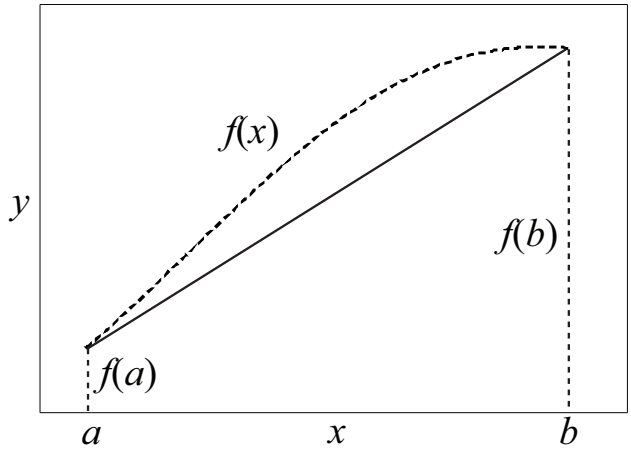

Рис. 4. Геометрическая иллюстрация простейшей формулы Ньютона – Котеса (основания трапеции — отрезки длиной $f(a)$ и $f(b)$, высота трапеции $(b–a)$

Итак, принципиально вычисление коэффициентов Котеса c_m^i несложно, но при больших m оно становится трудоемким. Впрочем, значения этих коэффициентов и не нужно вычислять, поскольку они табулированы. Пусть узлы интерполяции являются равноотстоящими с шагом h

$$x_0 = a,\ x_1 = a + h,\ x_2 = a + 2h, \ldots,\ x_m = a + mh = b.$$

Приведем фрагмент таблицы коэффициентов Котеса c_m^i при $m = 1, 2, \ldots, 6$.

Таблица 1

Коэффициенты Котеса c_m^i в случае равноотстоящих узлов

m	Коэффициенты Котеса c_m^i
1	$c_1^0 = c_1^1 = \dfrac{b-a}{2}$
2	$c_2^0 = c_2^2 = \dfrac{b-a}{6}, \quad c_2^1 = \dfrac{4(b-a)}{6}$
3	$c_3^0 = c_3^3 = \dfrac{b-a}{8}, \quad c_3^1 = c_3^2 = \dfrac{3(b-a)}{8}$
4	$c_4^0 = c_4^4 = \dfrac{7(b-a)}{90}, \quad c_4^1 = c_4^3 = \dfrac{16(b-a)}{45}, \quad c_4^2 = \dfrac{2(b-a)}{15}$
5	$c_5^0 = c_5^5 = \dfrac{19(b-a)}{288}, \quad c_5^1 = c_5^4 = \dfrac{25(b-a)}{96}, \quad c_5^2 = c_5^3 = \dfrac{25(b-a)}{144}$
6	$c_6^0 = c_6^6 = \dfrac{41(b-a)}{840}, \quad c_6^1 = c_6^5 = \dfrac{9(b-a)}{35}, \quad c_6^2 = c_6^4 = \dfrac{9(b-a)}{280}, \quad c_6^3 = \dfrac{34(b-a)}{105}$

Покажем, как пользоваться этой таблицей. Пусть функция $f(x)$ задана в трех точках: a, b и в точке $m = \dfrac{a+b}{2}$ — середине отрезка $[a, b]$. В таком случае, выбирая из строки $m = 2$ коэффициенты Котеса c_2^0, c_2^1, c_2^2, запишем определенный интеграл в виде

$$\int_a^b f(x)\,\mathrm{d}x \approx c_2^0 f(a) + c_2^1 f(m) + c_2^2 f(b) = \frac{b-a}{6} f(a) + \frac{4(b-a)}{6} f(m) + \frac{b-a}{6} f(b).$$

По формуле Ньютона – Котеса, которая является приближенным способом интегрирования, вычисляют определенный интеграл с неко-

торой погрешностью (кроме очевидного случая, когда подынтегральная функция является полиномом степени меньшей, чем порядок формулы Ньютона – Котеса, но тогда эта формула не нужна).

Погрешность R формулы Ньютона – Котеса — это модуль разности между точным значением интеграла и приближенным значением, получающимся при замене подынтегральной функции полиномом Лагранжа,

$$R_m = \left| \int_a^b f(x)\,\mathrm{d}x - \int_a^b L_m(x)\,\mathrm{d}x \right|.$$

Преобразуем это выражение и воспользуемся формулами (6), (7) оценки погрешности интерполяции

$$R_m = \left| \int_a^b \left[f(x) - L_m(x) \right] \mathrm{d}x \right| = \left| \int_a^b \frac{f^{(m+1)}(\xi)}{(m+1)!} (x - x_0)(x - x_1)\ldots(x - x_m)\,\mathrm{d}x \right| \le$$

$$\le \frac{M_{m+1}}{(m+1)!} \int_a^b \left| (x - x_0)(x - x_1)\ldots(x - x_m) \right| \mathrm{d}x. \qquad (22)$$

Напомним, что $\xi \in [a, b]$ — некоторая (неизвестная) точка отрезка интегрирования, $M_{m+1} = \max\limits_{x \in [a, b]} \left| f^{(m+1)}(x) \right|$. Поскольку найти ее нелегко, практическое значение оценки (22) ограниченно. Иногда применяют двойной пересчет интеграла с шагами h и $h/2$ и условно считают, что совпадающие десятичные знаки двух результатов являются верными цифрами.

Формула (22) непрактична, но из нее можно получить полезные выводы. Огрубим формулу (22) еще больше, учитывая, что каждый из сомножителей вида $|x - x_i|$ не превосходит $(b - a)$ — длины отрезка интегрирования

$$R_m \le \frac{M_{m+1}}{(m+1)!} \int_a^b (b - a)^{m+1}\,\mathrm{d}x = \frac{M_{m+1}(b - a)^{m+2}}{(m+1)!}.$$

Отсюда видно, что уменьшения погрешности формулы Ньютона – Котеса можно достичь двояко: увеличением ее порядка m и (или) сужением отрезка интегрирования (если M_{m+1} изменяется с ростом m незначительно). Но первый путь малопривлекателен: формула Нью-

тона – Котеса при большом m становится громоздкой и неудобна для использования. Удобнее воспользоваться второй возможностью: отрезок интегрирования разбить на узкие участки, на каждом из которых даже формула Ньютона – Котеса небольшого порядка m обеспечит достаточную точность. Таким путем мы придем к известным формулам численного интегрирования — формулам прямоугольников, трапеций и Симпсона.

Формула прямоугольников

Разобьем отрезок интегрирования $[a, b]$ на n равных элементарных отрезков с шагом $h = \frac{b-a}{n}$ точками $a_0 = a, a_1 = a + h, a_2 = a + 2h, \ldots,$ $a_n = a + nh = b$. На каждом элементарном отрезке $[a_{k-1}, a_k]$, $k = 1, 2, \ldots, n,$ аппроксимируем функцию $f(x)$ полиномом Лагранжа нулевой степени по значению $f(a_{k-1})$ на левом конце элементарного отрезка: $L_0(x) = f(a_{k-1}) = y_{k-1}$. Геометрически это означает замену криволинейной трапеции, ограниченной сверху графиком $f(x)$, ступенчатой фигурой (рис. 5).

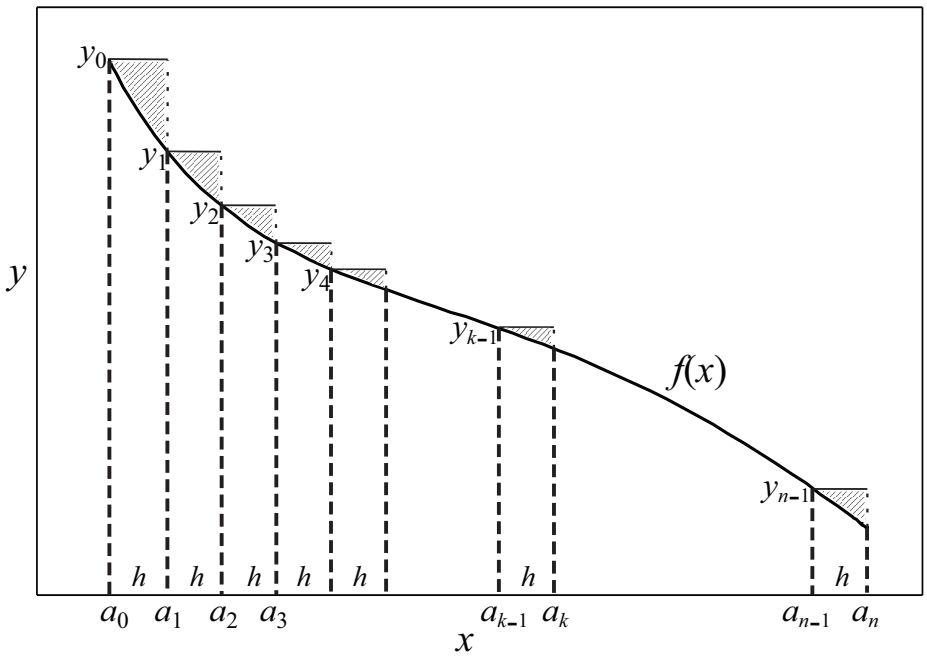

Рис. 5. Приближенное вычисление интеграла по формуле прямоугольников

Эта фигура состоит из прямоугольников с основанием h и высотами $y_i = f(a_i),\ i = 0, 1, n-1$. В таком случае интеграл $\int_a^b f(x)\,\mathrm{d}x$, численно равный площади криволинейной трапеции, приближенно получается как сумма площадей прямоугольников

$$\int_a^b f(x)\,\mathrm{d}x \approx y_0 h + y_1 h + \ldots + y_{n-1} h = h\left(y_0 + y_1 + \ldots + y_{n-1}\right). \qquad (23)$$

Получена формула прямоугольников приближенного вычисления интегралов.

Из рис. 5 видно, что погрешность этой формулы должна быть велика (она определяется суммарной площадью заштрихованных фигур). Для оценки погрешности воспользуемся формулой (22). На одном элементарном отрезке $[a_{k-1}, a_k]$

$$\left| \int_{a_{k-1}}^{a_k} f(x)\,\mathrm{d}x - \int_{a_{k-1}}^{a_k} L_0(x)\,\mathrm{d}x \right| \le \frac{M_1}{1!} \int_{a_{k-1}}^{a_k} |x - a_{k-1}|\,\mathrm{d}x =$$

$$= \frac{M_1}{1!} \int_{a_{k-1}}^{a_k} (x - a_{k-1})\,\mathrm{d}x = M_1 \frac{(x - a_{k-1})^2}{2} \Bigg|_{a_{k-1}}^{a_k} = \frac{M_1 h^2}{2},$$

где $M_1 = \max\limits_{[a,b]} |f'(x)|$. Область интегрирования $[a, b]$ содержит n элементарных отрезков, поэтому результирующая погрешность формулы прямоугольников

$$R_0 \le \frac{M_1 h^2}{2} n = \frac{M_1 h}{2}(b - a) = O(h), \qquad (24)$$

где мы учли, что $h = \frac{b-a}{n}$. Итак, погрешность линейно зависит от шага h разбиения области интегрирования на элементарные отрезки. Заметим, что при табличном задании $f(x)$ применить формулу (24), скорее всего, не удастся, поскольку значение M_1 взять неоткуда.

Формула трапеций

По-прежнему будем делить отрезок интегрирования $[a, b]$ на n элементарных отрезков точками $a_k = a + kh, \quad k = 1, 2, \ldots, n$, с шагом $h = \frac{b-a}{n}$. На каждом элементарном отрезке $[a_{k-1}, a_k]$ аппроксимируем функцию $f(x)$ полиномом Лагранжа $L_1(x)$ первой степени с узлами на концах отрезка в двух точках $x_0 = a_{k-1}$, $x_1 = a_k$. Поскольку графиком полинома первой степени является прямая, геометрическое значение этой аппроксимации заключается в замене каждой дуги кривой $[y_{k-1}, y_k]$ хордой (прямолинейным отрезком), стягивающей концы дуги. Уже здесь, не говоря о нижеследующей более точной формуле Симпсона, расхождение между графиком $f(x)$ и аппроксимирующей линией (здесь это ломаная, состоящая из прямолинейных звеньев) столь мало, что построить наглядную иллюстрацию, подобную рис. 5, затруднительно. Рассмотрим только один элементарный отрезок, преувеличивая расхождение между линиями (рис. 6).

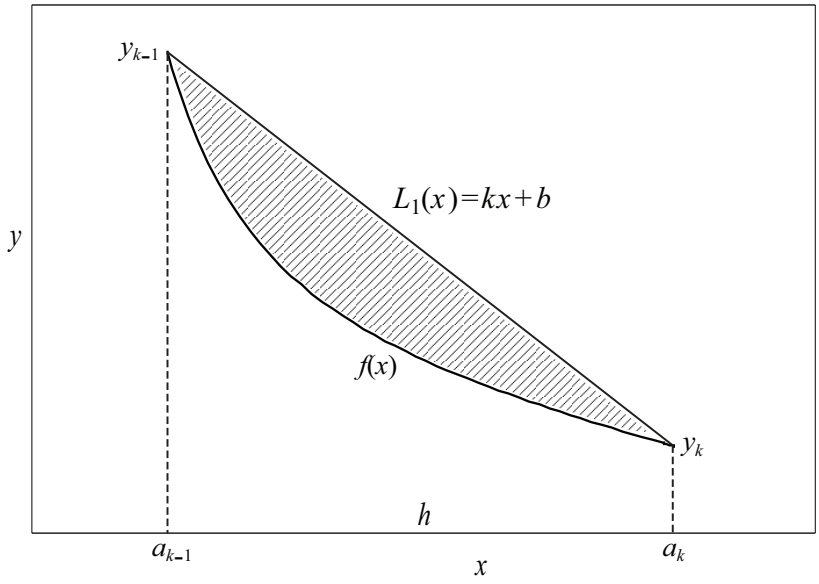

Рис. 6. Приближенное вычисление интеграла по формуле трапеций

На элементарном отрезке $[a_{k-1}, a_k]$ построена трапеция площадью $\frac{y_{k-1} + y_k}{2} h$, и из таких трапеций складывается фигура, приближающая

исходную криволинейную трапецию. В результате интеграл $\int_a^b f(x)\,\mathrm{d}x$, численно равный площади криволинейной трапеции, приближенно получается как сумма площадей трапеций:

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{y_0+y_1}{2}h + \frac{y_1+y_2}{2}h + \frac{y_2+y_3}{2}h + \ldots + \frac{y_{n-1}+y_n}{2}h = \\ = \frac{h}{2}\left[y_0 + 2\left(y_1 + y_2 + \ldots + y_{n-1}\right) + y_n\right]. \qquad (25)$$

Получена формула трапеций приближенного вычисления интегралов.

Совокупность сегментов, подобных заштрихованному на рис. 6, определяет погрешность формулы трапеций. Для оценки погрешности воспользуемся формулой (22). На одном элементарном отрезке $[a_{k-1}, a_k]$

$$\left|\int_{a_{k-1}}^{a_k} f(x)\,\mathrm{d}x - \int_{a_{k-1}}^{a_k} L_1(x)\,\mathrm{d}x\right| \le \frac{M_2}{2!}\int_{a_{k-1}}^{a_k}\left|(x-a_{k-1})(x-a_k)\right|\mathrm{d}x = \\ = \frac{M_2}{2!}\int_{a_{k-1}}^{a_k}\left|x^2 - x(a_{k-1}+a_k) + a_{k-1}a_k\right|\mathrm{d}x = \\ = -\frac{M_2}{2!}\int_{a_{k-1}}^{a_k}\left[x^2 - x(a_{k-1}+a_k) + a_{k-1}a_k\right]\mathrm{d}x = \frac{M_2 h^3}{12},$$

где $M_2 = \max\limits_{[a,b]}\left|f''(x)\right|$. Область интегрирования $[a,b]$ содержит n элементарных отрезков, поэтому результирующая погрешность формулы трапеций

$$R_1 \le \frac{M_2 h^3}{12} n = \frac{M_2 h^2}{12}(b-a) = O\left(h^2\right). \qquad (26)$$

Итак, погрешность квадратично зависит от шага h разбиения области интегрирования на элементарные отрезки, а поскольку при возведении малого числа в квадрат оно уменьшается, то для формулы трапеций погрешность будет меньше, чем для формулы прямоугольников (ср. с формулой (24)). По вышеназванной причине формула (26) зачастую оказывается не практичной.

Формула Симпсона

Продолжим разбивать отрезок интегрирования $[a, b]$ на n элементарных отрезков точками $a_k = a + kh, \quad k = 1, 2, \ldots, n$, с шагом $h = \frac{b-a}{n}$, но теперь количество разбиений пусть будет ч е т н ы м: $n = 2s$, где s — целое число. Рассмотрим два смежных элементарных отрезка с тремя узлами a_{2k-2}, a_{2k-1}, a_{2k} (рис. 7).

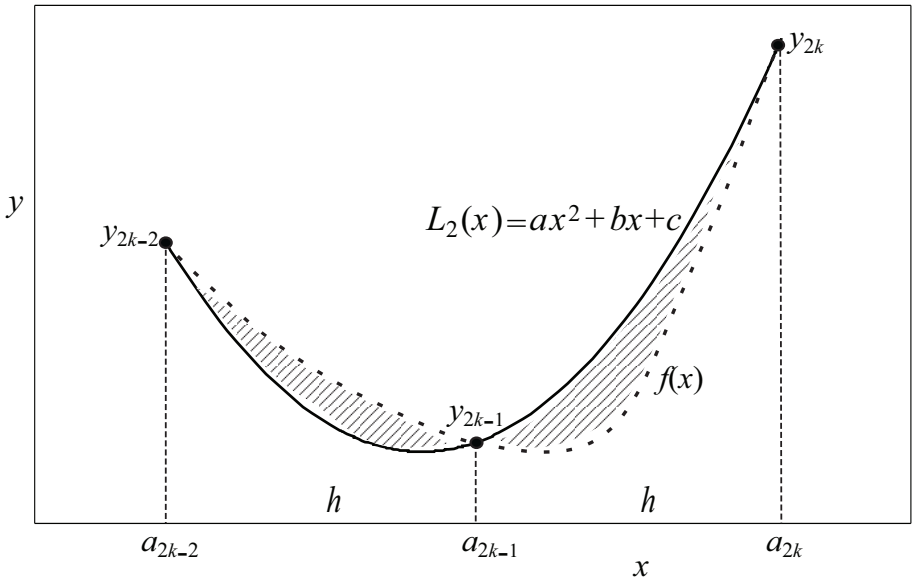

Рис. 7. Приближенное вычисление интеграла по формуле Симпсона

По значениям в трех узлах функция $f(x)$ аппроксимируется полиномом Лагранжа второй степени $L_2(x) = ax^2 + bx + c$, графиком которого является парабола. Дуги парабол очень близки к графику $f(x)$, поэтому расхождение между ними (штрихованные области) на рис. 7 преувеличено. По формуле Ньютона — Котеса для отрезка $[a_{2k-2}, a_{2k}]$, беря коэффициенты Котеса из табл. 1 при $m = 2$, получим

$$\int_{a_{2k-2}}^{a_{2k}} f(x)\,\mathrm{d}x \approx \int_{a_{2k-2}}^{a_{2k}} L_2(x)\,\mathrm{d}x = c_2^0 f(a_{2k-2}) + c_2^1 f(a_{2k-1}) + c_2^2 f(a_{2k}) =$$

$$= \frac{2h}{6} y_{2k-2} + 4 \cdot \frac{2h}{6} y_{2k-1} + \frac{2h}{6} y_{2k} = \frac{h}{3}(y_{2k-2} + y_{2k-1} + y_{2k}). \qquad (27)$$

Применяя формулу (27) к каждому отрезку $[a_{2k-2}, a_{2k}]$, $k = 1, 2, \ldots, s$, получим

$$\int_a^b f(x)\,\mathrm{d}x = \int_{a_0}^{a_2} f(x)\,\mathrm{d}x + \int_{a_2}^{a_4} f(x)\,\mathrm{d}x + \ldots + \int_{a_{2s-2}}^{a_{2s}} f(x)\,\mathrm{d}x \approx$$

$$\approx \frac{h}{3}\left[(y_0 + 4y_1 + y_2) + (y_2 + 4y_3 + y_4) + (y_4 + 4y_5 + y_6) + \ldots + (y_{2s-2} + 4y_{2s-1} + y_{2s})\right].$$

Итак,

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{h}{3}\left[(y_0 + y_{2s}) + 2(y_2 + y_4 + \ldots + y_{2s-2}) + 4(y_1 + y_3 + \ldots + y_{2s-1})\right]. \quad (28)$$

Получена формула Симпсона (1710–1761) приближенного вычисления определенного интеграла.

Оценим погрешность с помощью формулы (22). Для одного сдвоенного элементарного отрезка $[a_{2k-2}, a_{2k}]$

$$\left|\int_{a_{2k-2}}^{a_{2k}} f(x)\,\mathrm{d}x - \int_{a_{2k-2}}^{a_{2k}} L_2(x)\,\mathrm{d}x\right| \le \frac{M_3}{3!}\int_{a_{2k-2}}^{a_{2k}} \left|(x - a_{2k-2})(x - a_{2k-1})(x - a_{2k})\right|\mathrm{d}x = \frac{M_3 h^4}{96},$$

где $M_3 = \max\limits_{[a,b]} |f'''(x)|$. Область интегрирования $[a, b]$ содержит $n/2$ сдвоенных элементарных отрезков, поэтому результирующая погрешность формулы Симпсона

$$R_2 \le \frac{M_3 h^4}{96}\frac{n}{2} = \frac{M_3 h^3}{192}(b - a) = O(h^3). \quad (29)$$

Формула Симпсона оказывается точнее формулы трапеций, не говоря уже о формуле прямоугольников. Понятно, что при табличном задании подынтегральной функции воспользоваться формулой (29) для оценки погрешности численного интегрирования практически нереально, поскольку значение M_3 взять неоткуда.

Пример. Вычислить интеграл $I = \int_1^2 \frac{\mathrm{d}x}{x}$ по формуле Ньютона – Котеса при $m = 4$, а также по формулам прямоугольников, трапеций и Симпсона при $n = 4$; сравнить результаты с точным значением интеграла.

Решение. Будем вести вычисления с пятью десятичными знаками. Для аналитически заданной подынтегральной функции $f(x)=1/x$ оценить погрешность численного интегрирования по формулам (22), (24), (26), (29) возможно, но излишне, поскольку интеграл точно вычисляется по формуле Ньютона – Лейбница, с чего мы и начнем:

$$I=\int_1^2 \frac{\mathrm{d}x}{x}=\ln x\Big|_1^2=\ln 2=\mathbf{0.693\,15}\ldots$$ (бесконечная непериодическая дробь).

Интерполяционный полином Лагранжа степени $m=4$ в формуле Ньютона – Котеса используется, если на отрезке интегрирования есть пять равноудаленных узлов. В данном случае шаг (расстояние между узлами) $h=\dfrac{b-a}{4}=\dfrac{2-1}{4}=0.25$.

Для подынтегральной функции $f(x)=1/x$ составим следующую таблицу (табл. 2).

Таблица 2

Значения функции $f(x)=1/x$ в узлах

k	$x_k=x_0+kh$	$y_k=1/x_k$
0	$x_0=1$	$y_0=1$
1	$x_1=x_0+h=1.25=5/4$	$y_1=1\Big/\dfrac{5}{4}=4/5=0.800\,00$
2	$x_2=x_1+h=1.5=3/2$	$y_2=1\Big/\dfrac{3}{2}=2/3=0.666\,67$
3	$x_3=x_2+h=1.75=7/4$	$y_3=1\Big/\dfrac{7}{4}=4/7=0.571\,43$
4	$x_4=x_3+h=2$	$y_4=1/2=0.500\,00$

Приближенное значение интеграла по формуле Ньютона – Котеса при $m=4$

$$I_{\text{Cotes}}=\sum_{i=0}^{4}c_4^i y_i=c_4^0y_0+c_4^1y_1+c_4^2y_2+c_4^3y_3+c_4^4y_4=$$

$$=\frac{7}{90}\cdot 1.000\,\mathbf{0}0+\frac{16}{45}\cdot 0.800\,00+\frac{2}{15}\cdot 0.666\,67+\frac{16}{45}\cdot 0.571\,43+$$

$$+\frac{7}{90}\cdot 0.500\,00=\mathbf{0.693\,18}.$$

Сравнивая результат с точным, видим, что формула Ньютона – Котеса дает четыре верных десятичных знака.

Приближенное значение интеграла по формуле прямоугольников (23)

$$I_{\text{прям}} = h\left[y_0 + y_1 + y_2 + y_3\right] =$$

$$= 0{,}25 \cdot \left[1.000\,00 + 0.800\,00 + 0.666\,67 + 0.571\,43\right] = \mathbf{0.759\,53}.$$

Ни одного верного десятичного знака!

Приближенное значение интеграла по формуле трапеций (25)

$$I_{\text{трап}} = \frac{h}{2}\left[y_0 + 2\left(y_1 + y_2 + y_3\right) + y_4\right] = \mathbf{0.697\,02}\,.$$

Два верных знака после запятой.

Приближенное значение интеграла по формуле Симпсона (28)

$$I_{\text{Simp}} = \frac{h}{3}\left[\left(y_0 + y_4\right) + 4\left(y_1 + y_3\right) + 2y_2\right] = \mathbf{0.693\,25}\,.$$

Три верных знака после запятой.

Итак, точнее всех, как и можно было ожидать, оказалась формула Ньютона – Котеса. Ее недостатки — громоздкость и необходимость привлечения специальных таблиц (коэффициентов Котеса). Формулы, являющиеся упрощенными следствиями из формулы Ньютона – Котеса, дают не столь точные результаты. При этом формула Симпсона, будучи гораздо проще формулы Ньютона – Котеса, успешно с ней конкурирует по точности. Формула прямоугольников столь груба (при взятом нами шаге h), что вряд ли применима на практике.

Видимо, формулы трапеций и Симпсона являются самыми «практичными», сочетая простоту и удовлетворительную точность.

Заметим еще раз, что при табличном задании подынтегральной функции попытка оценки погрешности численного интегрирования по формулам (22), (24), (26), (29) наталкивается на непреодолимые трудности.

Приближенное решение обыкновенных дифференциальных уравнений

Напомним некоторые нужные для дальнейшего понятия. **Решением** дифференциального уравнения (ДУ) I порядка $y' = f(x, y)$, разрешенного относительно производной, называется функция $y = \varphi(x)$, которая при подстановке в уравнение обращает его в тождество:

$$\varphi'(x) \equiv f(x, \varphi(x)).$$

Задача Коши

$$\left.\begin{aligned} y' &= f(x, y), \\ y(x_0) &= y_0 \end{aligned}\right\} \tag{67}$$

состоит в нахождении частного решения ДУ, удовлетворяющего начальному условию $y(x_0) = y_0$. Геометрический смысл задачи Коши: найти такую интегральную кривую (график решения), которая проходит через заданную начальную точку $M_0(x_0, y_0)$.

Аналитическое решение задачи Коши (или решение в квадратурах) заключается в получении частного решения путем выполнения конечного числа операций дифференцирования, интегрирования и арифметических действий (сложение, вычитание, умножение, деление). К сожалению, круг задач, решаемых в квадратурах, крайне узок, поэтому актуальна задача приближенного, численного интегрирования ДУ.

Решить задачу Коши численно — значит для заданной последовательности значений аргумента (узлов) $x_0, x_1, \ldots, x_n$ и числа y_0 (значение искомой функции в начальном узле x_0), не находя самого́ решения $y = \varphi(x)$, приближенно вычислить значения $y_1, y_2, \ldots, y_n$ этого решения в остальных узлах. Численное решение задачи Коши позво-

ляет вместо отыскания точного решения $y=\varphi(x)$ в виде формулы получить таблицу значений

x_i	x_0	x_1	x_2	...	x_n
$\varphi(x_i)$	$y_0=\varphi(x_0)$	y_1	y_2	...	y_n

этой функции[18]. Рассмотрим некоторые способы численного интегрирования ДУ.

Метод ломаных Эйлера

Метод (ломаных) Эйлера (1707–1783) основан на кусочной замене искомой функции полиномом первой степени, т. е. на линейной интерполяции. Впрочем, точнее было бы говорить о линейной экстраполяции, т. к. речь идет о нахождении значений функции $y=\varphi(x)$ в соседних узлах, а не между узлами.

Выбрав малый шаг h, построим систему равноотстоящих узлов $x_0, x_1, x_2, \ldots$, где $x_k = x_0 + kh$ (рис. 18).

В начальной точке $M_0(x_0, y_0)$ проведем прямую с угловым коэффициентом $k_0 = y'(x_0)$ $(= f(x_0, y_0)$ в силу дифференциального уравнения (67). Эта прямая является касательной к (неизвестной) интегральной кривой — графику искомого решения $y=\varphi(x)$ задачи Коши. Уравнение касательной

$$y - y_0 = k_0(x - x_0) \text{ или } y = y_0 + f(x_0, y_0)\cdot(x - x_0).$$

В качестве приближенного решения задачи Коши (67) в узле x_1 примем ординату y_1 точки пересечения касательной с вертикальной прямой $x = x_1$:

$$y_1 = y_0 + f(x_0, y_0)\cdot(x_1 - x_0) = y_0 + f(x_0, y_0)\cdot h.$$

Через точку $M_1(x_1, y_1)$ проведем прямую с угловым коэффициентом $k_1 = y'(x_1)$ $(= f(x_1, y_1)$ в силу ДУ (67). Ее уравнение

$$y = y_1 + f(x_1, y_1)\cdot(x - x_1). \qquad (68)$$

[18] Потом можно получить и приближенное задание функции формулой, например, аппроксимируя ее полиномом Лагранжа.

Эта прямая уже не является касательной к интегральной кривой $y=\varphi(x)$, т. к. точка $M_1(x_1, y_1)$ фиктивная, она не лежит на интегральной кривой. Пересечение прямой (68) с вертикалью $x=x_2$ — точка M_2 — имеет ординату $y_2 = y_1 + f(x_1, y_1) \cdot (x_2 - x_1) = y_1 + f(x_1, y_1) \cdot h$.

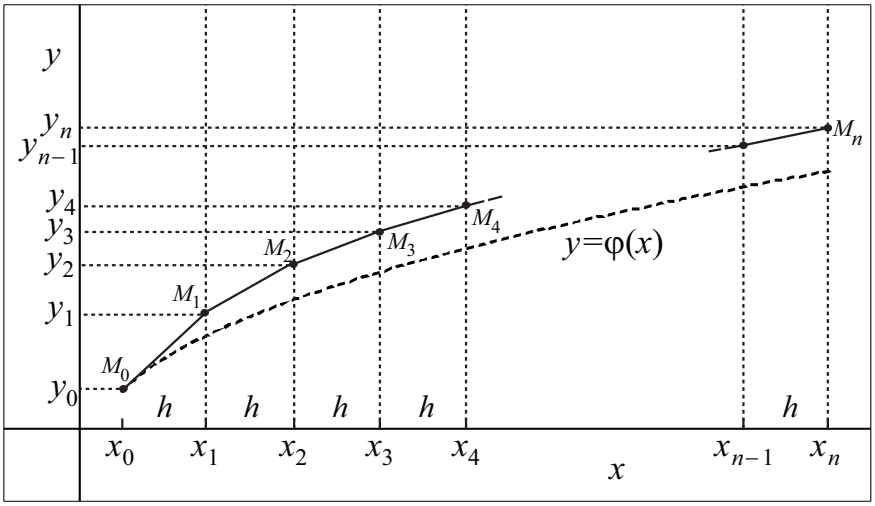

Рис. 18. Графическая иллюстрация метода Эйлера[19]

Через точку $M_2(x_2, y_2)$ проведем прямую с угловым коэффициентом $k_2 = y'(x_2) = f(x_2, y_2)$ до пересечения с вертикалью $x = x_3$. Ордината точки пересечения $y_3 = y_2 + f(x_2, y_2) \cdot h$. Точка $M_3(x_3, y_3)$ находится дальше по вертикали от интегральной кривой $y=\varphi(x)$, чем M_2 (см. рис. 18). Погрешность метода с каждым следующим шагом увеличивается.

Таким образом получаем ломаную Эйлера, составленную из отрезков прямых,

$$y = y_k + f(x_k, y_k) \cdot (x - x_k), \; k = 0, 1, 2, \ldots$$

Ордината каждой последующей угловой точки ломаной

$$y_{k+1} = y_k + h \cdot f(x_k, y_k). \qquad (69)$$

[19] Ломаная Эйлера (сплошная линия) приближенно представляет гипотетический график решения задачи Коши (штриховая линия). Последний график неизвестен, поскольку задача Коши не решена в квадратурах. Обе линии исходят из начальной точки $M_0(x_0, y_0)$; расхождение между ними постепенно возрастает.

Метод ломаных Эйлера груб и дает удовлетворительную точность лишь при малом шаге h. Действительно, разложим точное решение ДУ $y' = f(x, y)$ в ряд Тейлора в окрестности узла x_k:

$$y(x_k + h) = y(x_{k+1}) = y(x_k) + y'(x_k) \cdot h + O(h^2) =$$
$$= y(x_k) + f(x_k, y_k) \cdot h + O(h^2).$$

Сравнивая эту формулу с формулой (69), получаем, что погрешность последней того же порядка малости, что и h^2.

Пример. Решить задачу Коши

$$\left.\begin{aligned} y' &= -y, \\ y(0) &= 1 \end{aligned}\right\}$$

с шагом $h = 0{,}2$ на отрезке $[0, 1]$ методом Эйлера; сравнить приближенное решение с точным.

Решение. Начнем с решения в квадратурах. Дифференциальное уравнение

$$y' = \frac{dy}{dx} = -y$$

является уравнением с разделяющимися переменными. Разделим их

$$\frac{dy}{y} = -dx,$$

проинтегрируем

$$\int \frac{dy}{y} = -\int dx; \quad \ln|y| = -x + C; \quad y = e^{C-x}.$$

Общее решение ДУ найдено. Накладываем на него начальное условие

$$y(0) = e^{C-0} = 1; \quad C = 0.$$

Итак, частное решение ДУ (решение задачи Коши) $y = e^{-x}$.

Теперь применим приближенный метод Эйлера. В обозначениях выражения (67)

$$f(x, y) = -y, \quad x_0 = 0; \quad y_0 = 1.$$

В таком случае в силу формулы (69),

$$y_{k+1} = y_k + h \cdot f(x_k, y_k) = y_k + 0.2 \cdot f(x_k, y_k) = y_k - 0.2y_k = 0.8y_k,$$
$$k = 0, 1, 2, \mathbf{3}, 4, 5.$$

Составим сопоставительную таблицу значений y_k приближенного решения по Эйлеру и значений точного решения $y_k^* = e^{-x_k}$:

k	0	1	2	3	4	5
$x_k = x_0 + hk$	0	0.2	0.4	0.6	0.8	1
y_k	1	0.8	0.64	0.512	0.4096	0.327 68
y_{k^*}	1	0.8187	0.6703	0.5488	0.4493	0.3679

Грубое согласие между точными и приближенными результатами есть, но оно постепенно ухудшается.

Выводы по методу Эйлера следующие.

1. Расчетные формулы метода Эйлера:

$$y_{k+1} = y_k + \Delta y_k, \text{ где } \Delta y_k = h \cdot f(x_k, y_k), \quad k = 0, 1, 2, \ldots$$

2. Метод Эйлера — представитель одношаговых приближенных методов, в которых решение в $(k+1)$-м узле получается на основе решения только в одном предыдущем k-м узле. Тем самым информация о более ранних уже вычисленных значениях игнорируется. «Расточительный» подход к получаемым результатам оборачивается повышенным объемом вычислений. Одношаговые методы не самые экономичные в этом смысле.

3. Как и в любом одношаговом методе, начиная со второго шага исходное значение y_k в формуле $y_{k+1} = y_k + \Delta y_k$ самó является приближенным, т. е. погрешность на каждом последующем шаге систематически возрастает.

4. Оценка погрешности метода затруднительна. Часто пользуются эмпирическим правилом двойного пересчета (половинного шага): дважды проходят заданный отрезок интегрирования ДУ с шагами h и $h/2$. Совпадение соответствующих десятичных знаков в полученных результатах дает основание считать эти знаки верными.

5. Уменьшение h повышает точность вычислений, но резко увеличивает их объем. В целом метод ломаных Эйлера применим только для грубой прикидки.

Метод последовательного дифференцирования

К методу Эйлера можно прийти и без геометрических построений. Разложим искомое решение задачи Коши

$$\left.\begin{aligned} y' &= f(x,y), \\ y(x_0) &= y_0 \end{aligned}\right\}$$

в ряд Тейлора в окрестности начальной точки x_0

$$y(x) = \underbrace{y(x_0)}_{=y_0\,(\text{н.у.})} + \underbrace{y'(x_0)}_{\substack{=f(x_0,y_0) \\ \text{в силу ДУ}}} \cdot (x - x_0) + \frac{y''(x_0)}{2!}(x - x_0)^2 + \ldots \qquad (70)$$

Ограничиваясь в разложении первыми двумя (линейными по x) слагаемыми и полагая $x = x_1$, снова получим формулу (69) метода Эйлера

$$y(x_1) = y_1 = y_0 + f(x_0, y_0) \cdot (x_1 - x_0) = y_0 + h \cdot f(x_0, y_0).$$

Учтем теперь еще одно квадратичное по x слагаемое в формуле (70). Для этого потребуется вычислить $y''(x_0)$. Продифференцируем по x обе части ДУ $y' = f(x,y)$:

$$(y')' = y''(x) = \frac{\mathrm{d}}{\mathrm{d}x} f(x, y(x)) = f'_x(x,y) + f'_y(x,y) \cdot \underbrace{\frac{\mathrm{d}y}{\mathrm{d}x}}_{\substack{=f(x,y) \\ \text{в силу ДУ}}} = f'_x + f'_y \cdot f.$$

Полагая в выражении (70) $x = x_1$, получим во втором порядке разложения

$$y(x_1) = y_1 = y_0 + h \cdot f(x_0, y_0) + \frac{1}{2}\left[f'_x(x,y) + f'_y(x,y) \cdot f(x,y) \right]\Big|_{\substack{x = x_0 \\ y = y_0}} \cdot h^2.$$

Первые два слагаемых в правой части соответствуют методу Эйлера, а третье — поправка к нему. Для произвольного узла

$$y_{k+1} = y_k + h \cdot f(x_k, y_k) + \frac{1}{2}\left[f'_x + f'_y \cdot f \right]\Big|_{\substack{x = x_k \\ y = y_k}} \cdot h^2. \qquad (71)$$

Удерживая в ряду Тейлора (70) больше слагаемых, можно было бы получить сколь угодно точные формулы приближенного решения задачи Коши.

В этом состоит метод последовательного дифференцирования, который еще называют методом разложения решения в степенной ряд.

Неудобство разложения решения ДУ в степенной ряд связано с тем, что в расчетные формулы, наряду с $f(x,y)$, входят ее частные производные. В выведенную формулу (71), включающую слагаемые до h^2 включительно, входят f'_x, f'_y; в формулу с h^3 войдут вторые частные производные, и т. д. Но вычисление частных производных трудно автоматизировать — это задача не для стандартно применяемых языков программирования[20]. Поэтому частные производные пришлось бы искать вручную и конструировать с их помощью громоздких формул.

Альтернативой методу последовательного дифференцирования является метод Рунге – Кутты, лишенный отмеченного недостатка.

Метод Рунге — Кутты

Идея, предложенная Рунге (1856–1927) и Куттой (1867–1944), заключается в том, чтобы при численном решении задачи Коши (67) не использовать в расчетных формулах частные производные функции $f(x,y)$; использовать только ее саму, зато вычислять на каждом шаге ее значения в нескольких точках[21].

Проиллюстрируем это на примере одного из возможных методов Рунге – Кутты II порядка. Из определения производной

$$y'(x_k) = \lim_{\Delta x \to 0} \frac{\Delta y(x_k)}{\Delta x} = \lim_{x_{k+1} \to x_k} \frac{y(x_{k+1}) - y(x_k)}{x_{k+1} - x_k} = \lim_{h \to 0} \frac{y_{k+1} - y_k}{h} \approx \frac{y_{k+1} - y_k}{h}. \quad (72)$$

В финальное выражение входят значения функции y в двух точках, а производной уже нет. Подставим это приближенное выражение для производной в решаемое ДУ $y' = f(x,y)$, беря значение правой части в k-м узле

[20] Разве что для языков сверхвысокого уровня — Mathematica, Maple, Matlab.

[21] В сущности эта же идея применялась нами в методах хорд и секущих численного решения уравнений — см. формулу (37).

$$\frac{y_{k+1}-y_k}{h}=f\left(x_k,y_k\right).$$

Отсюда $y_{k+1}=y_k+h\cdot f\left(x_k,y_k\right)$, и мы снова получили метод Эйлера (69)! Но поскольку для аппроксимации производной y' взяты две точки, то и для правой части ДУ $f\left(x,y\right)$ уместно привлечь две точки:

$$\frac{1}{2}\left[f\left(x_k,y_k\right)+f\left(x_{k+1},y_{k+1}\right)\right]. \tag{73}$$

Совмещая выражения (72) и (73)

$$\frac{y_{k+1}-y_k}{h}=\frac{1}{2}\left[f\left(x_k,y_k\right)+f\left(x_{k+1},y_{k+1}\right)\right],$$

после преобразования получим

$$y_{k+1}=y_k+\frac{h}{2}\left[f\left(x_k,y_k\right)+f\left(x_{k+1},y_{k+1}\right)\right]. \tag{74}$$

Искомой величиной в уравнении (74) является y_{k+1}, входящей в обе части уравнения. Решать это уравнение можно методом итераций, беря в качестве начального приближения то значение $y_{k+1}=y_k+h\cdot f\left(x_k,y_k\right)$, которое получается в методе Эйлера. Тогда

$$y_{k+1}=y_k+\frac{h}{2}\left[f\left(x_k,y_k\right)+f\left(x_k+h,\ \underbrace{y_k+h\cdot f\left(x_k,y_k\right)}_{\text{Эйлер}}\right)\right] \tag{75}$$

или, для придания этой формуле стандартного вида,

$$y_{k+1}=y_k+\frac{1}{2}\left(r_1+r_2\right), \tag{76}$$

где

$$r_1=h\cdot f\left(x_k,y_k\right), \tag{77}$$

$$r_2=h\cdot f\left(x_k+h,y_k+r_1\right). \tag{78}$$

Формулы (76)–(78) представляют метод Рунге – Кутты II порядка[22]. Применяют их в такой последовательности: сначала находят r_1 (77), его подставляют в r_2 (78); в заключение вычисляют y_{k+1} (76).

[22] Формально порядок метода можно связать с количеством величин r_i, или, что то же, с количеством точек, в которых вычисляется значение функции $f(x, y)$.

Покажем, что эти формулы с точностью до h^2 включительно согласуются с формулой (71), полученной другим способом. Для этого выражение $f\left(x_k + h, y_k + h \cdot f\left(x_k, y_k\right)\right)$, входящее в формулу (75), разложим в ряд по степеням h до линейных слагаемых включительно[23]:

$$f\left(x_k + h, y_k + h \cdot f\left(x_k, y_k\right)\right) \approx$$
$$\approx f\left(x_k, y_k\right) + f'_x\left(x_k, y_k\right) \cdot h + f'_y\left(x_k, y_k\right) \cdot h f\left(\boldsymbol{x}_k, y_k\right).$$

Подставим преобразованное выражение в формулу (75), откуда оно было извлечено:

$$y_{k+1} \approx y_k + \frac{h}{2} \cdot f\left(x_k, y_k\right) +$$
$$+ \frac{h}{2}\left[f\left(x_k, y_k\right) + f'_x\left(x_k, y_k\right) \cdot h + f'_y\left(x_k, y_k\right) \cdot h f\left(\boldsymbol{x}_k, y_k\right)\right] =$$
$$= y_k + h \cdot f\left(x_k, y_k\right) + \frac{h^2}{2}\left[f'_x\left(x_k, y_k\right) + f'_y\left(x_k, y_k\right) \cdot f\left(x_k, y_k\right)\right].$$

Это точно совпадает с формулой (71), выведенной методом последовательного дифференцирования. Однако формулы (76)–(78) не требуют вычисления частных производных, поэтому они удобнее.

Подробно ознакомившись с методом Рунге–Кутты II порядка, заметим, что чаще используются родственные более точные (но и более громоздкие) методы Рунге – Кутты высших порядков.

Один из самых известных — метод Рунге–Кутты IV порядка, часто без уточнений называемый просто методом Рунге – Кутты,

$$y_{k+1} = y_k + \frac{1}{6}\left(r_1 + 2r_2 + 2r_3 + r_4\right), \tag{79}$$

где

$$r_1 = h \cdot f\left(x_k, y_k\right),$$
$$r_2 = h \cdot f\left(x_k + \frac{h}{2}, y_k + \frac{r_1}{2}\right),$$

[23] Вспомним, что ряд Тейлора для функции двух переменных имеет вид $f(x,y) = f\left(x_0, y_0\right) + f'_x\left(x_0, y_0\right) \cdot \left(x - x_0\right) + f'_y\left(x_0, y_0\right) \cdot \left(y - y_0\right) + \ldots$

$$r_3 = h \cdot f\left(x_k + \frac{h}{2}, y_k + \frac{r_2}{2} \right),$$

$$r_4 = h \cdot f\left(x_k + h, y_k + r_3 \right),$$

причем сначала последовательно вычисляются r_1, r_2, r_3, r_4, а затем — y_{k+1} (79).

Пример. Решить задачу Коши

$$\left.\begin{array}{l} y' = y(1-x), \\ y(0) = 1 \end{array}\right\} \tag{80}$$

методом Рунге – Кутты на отрезке $[0; 0.5]$ с шагом $h = 0.05$. Сравнить полученное решение с точным и решением по методу Эйлера.

Решение. Начнем с нахождения точного решения. ДУ относится к уравнениям с разделяющимися переменными:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = y(1-x); \quad \frac{\mathrm{d}y}{y} = (1-x)\mathrm{d}x; \quad \int \frac{\mathrm{d}y}{y} = \int (1-x)\mathrm{d}x; \quad \ln|y| = x - x^2/2 + \ln C.$$

Итак, общее решение ДУ (80):

$$y(x) = Ce^{x - x^2/2}.$$

Применим начальное условие:

$$y(0) = Ce^0 = 1; \quad C = 1.$$

Итак, получено точное решение $y(x) = e^{x - x^2/2}$ задачи Коши.

Подробно покажем первый (из десяти!) этап приближенного решения методом Рунге – Кутты. Правая часть ДУ $f(x, y) = y(1-x)$; $x_0 = 0$, $y_0 = 1$. Вычислим величины r_i:

$$r_1 = h\, f(x_0, y_0) = 0.05 \cdot y_0 (1 - x_0) = 0.05 \cdot 1 \cdot (1 - 0) = 0.05;$$

$$r_2 = h\, f\left(x_0 + \frac{h}{2}, y_0 + \frac{r_1}{2} \right) = 0.05 \left(y_0 + \frac{r_1}{2} \right) \left(1 - \left(x_0 + \frac{h}{2} \right) \right) =$$

$$0.05 \cdot \left(1 + \frac{0.05}{2} \right) \cdot \left(1 - \left(0 + \frac{0.05}{2} \right) \right) = 0.049\,97;$$

$$r_3 = h\,f\left(x_0 + \frac{h}{2}, y_0 + \frac{r_2}{2}\right) = 0.05\left(y_0 + \frac{r_2}{2}\right)\left(1 - \left(x_0 + \frac{h}{2}\right)\right) =$$

$$= 0.05 \cdot \left(\mathbf{1} + \frac{0.049\,97}{2}\right) \cdot \left(1 - \left(0 + \frac{0.05}{2}\right)\right) = 0.049\,97;$$

$$r_4 = h\,f\left(x_0 + h, y_0 + r_3\right) = 0{,}05\left(y_0 + r_3\right)\left(1 - \left(x_0 + h\right)\right) =$$

$$= 0.05 \cdot \left(1 + 0.049\,97\right) \cdot \left(\mathbf{1} - \left(0 + 0.05\right)\right) = 0.049\,87.$$

Тогда

$$y_1 = y_0 + \frac{1}{6}\left(r_1 + 2r_2 + 2r_3 + r_4\right) =$$

$$= 1 + \frac{1}{6} \cdot \left(0.05 + 2 \cdot 0.049\,97 + 2 \cdot 0.049\,97 + 0.0\mathbf{4}9\,87\right) = 1.049\,958.$$

Сравнение решений, получаемых разными методами, представлено в виде табл. 4.

Таблица 4

Решения, полученные разными методами

x	y		
	Эйлер	Рунге – Кутта IV порядка	точное решение $y = e^{x - x^2/2}$
0	1	1	1
0.05	1.05*	1.049 958	1.049 958
0.1	1.099 875**	1.099 659	1.099 659
0.15	1.149 369	1.148 837	1.148 837
. . . . .			
0.45	1.423 065	1.417 295	1.417 295
0.5	1.462 199	1.454 991	1.454 991

Примечания: *, ** — все вычисления производятся методом Эйлера.

* $y_1 = y_0 + h f\left(x_0, y_0\right) = 1 + h y_0\left(1 - x_0\right) = 1 + 0.05 \cdot 1 \cdot \left(1 - 0\right) = 1.05.$

** $y_2 = y_1 + h f\left(x_1, y_1\right) = 1.05 + 0.05\, y_1\left(1 - x_1\right) = 1.05 + 0.05 \cdot 1.05 \cdot \left(1 - 0.05\right) = \mathbf{1}.099\,875.$

Отметим превосходное согласие результатов в двух последних столбцах.

Итоги по численному решению ДУ.

1. Метод Рунге – Кутты, сравнимый по точности с методом разложения в степенной ряд, лучше поддается автоматизации на ЭВМ, поскольку не требует вычисления частных производных. Метод Эйлера уступает по точности этим методам.

2. Все рассмотренные методы решения ДУ — метод Эйлера, метод разложения в степенной ряд, метод Рунге – Кутты — являются одноша-

говыми. Напомним, что это означает построение y_{k+1} на основе только y_k с игнорированием более ранних предшествующих результатов.

3. Все одношаговые методы имеют проблемы с оценкой погрешности результатов. На практике применяется эмпирическое правило двойного пересчета (половинного шага).

4. Все одношаговые методы сопряжены с избыточными вычислениями, объем которых можно существенно уменьшить при более рачительном использовании уже полученных результатов.

5. Альтернативой одношаговым являются многошаговые методы интегрирования ДУ — семейство методов Адамса [24]. В этих методах для вычисления y_{k+1} используется несколько значений приближенного решения на предыдущих шагах: y_k, y_{k-1}, y_{k-2},… Поскольку к моменту вычисления y_{k+1} они уже найдены, можно избежать многочисленных вычислений значений $f(x,y)$. Но, чтобы метод Адамса мог стартовать, первые значения y_0, y_1,… все-таки приходится находить одношаговыми методами. В целом методы Адамса в несколько раз менее трудоемки, чем метод Рунге – Кутты.