§ 2. Решение уравнений с одной переменной.

Решение нелинейных уравнений с одной переменной (или систем таких уравнений) представляет собой одну из важных задач прикладного анализа. Эта проблема возникает в многочисленных и разнообразных разделах физики, механики, математики и техники.

Примером может служить проблема определения по заданным законам движения материальной точки того момента времени, когда эта частица попадает в определенную точку физического пространства.

В газовой динамике мы встречаемся с задачей определения состояния газа после прохождения по нему ударной волны и т.д.

В общем случае нелинейное уравнение можно записать в виде:

$$F(x) = 0, \tag{2.1}$$

где функция $F(x)$ определена и непрерывна на некотором отрезке $[a, b]$.

Всякое число $\xi \in [a, b]$, обращающее функцию $F(x)$ в нуль, т.е. такое, при котором $F(\xi) = 0$, называется ***корнем уравнения*** (2.1) или ***нулем функции*** $F(x)$.

Число ξ называется ***корнем k-й кратности***, если при $x = \xi$ вместе с функцией $F(x)$ равны нулю ее производные до $(k-1)$ порядка включительно – $F(\xi) = F'(\xi) = \ldots = F^{(k-1)}(\xi) = 0$.

Однократный корень называется ***простым***.

Два уравнения $F(x) = 0$ и $G(x) = 0$ называются ***равносильными*** (эквивалентными), если всякое решение одного из них является решением и для другого, т.е. множества решений этих уравнений совпадают, при условии обязательного совпадения областей определения этих функций.

Нелинейные уравнения с одним неизвестным подразделяются на ***алгебраические*** и ***трансцендентные***.

Уравнение (2.1) называется ***алгебраическим***, если выражение $F(x)$ является алгебраической функцией.

Путем алгебраических преобразований из всякого алгебраического уравнения можно получить уравнение в канонической форме:

$$P_n(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0 = 0, \tag{2.2}$$

где a_0, a_1, …, a_n – коэффициенты уравнения, а x – неизвестное.

Показатель n называют степенью алгебраического уравнения.

Известно (из курса высшей алгебры), что всякое алгебраическое уравнение вида (2.2) имеет ровно n комплексных корней.

Существует ряд специальных численных методов решения уравнений (2.2): Берстоу-Хичкока, Берстоу-Ньютона, схема Горнера и т.д.

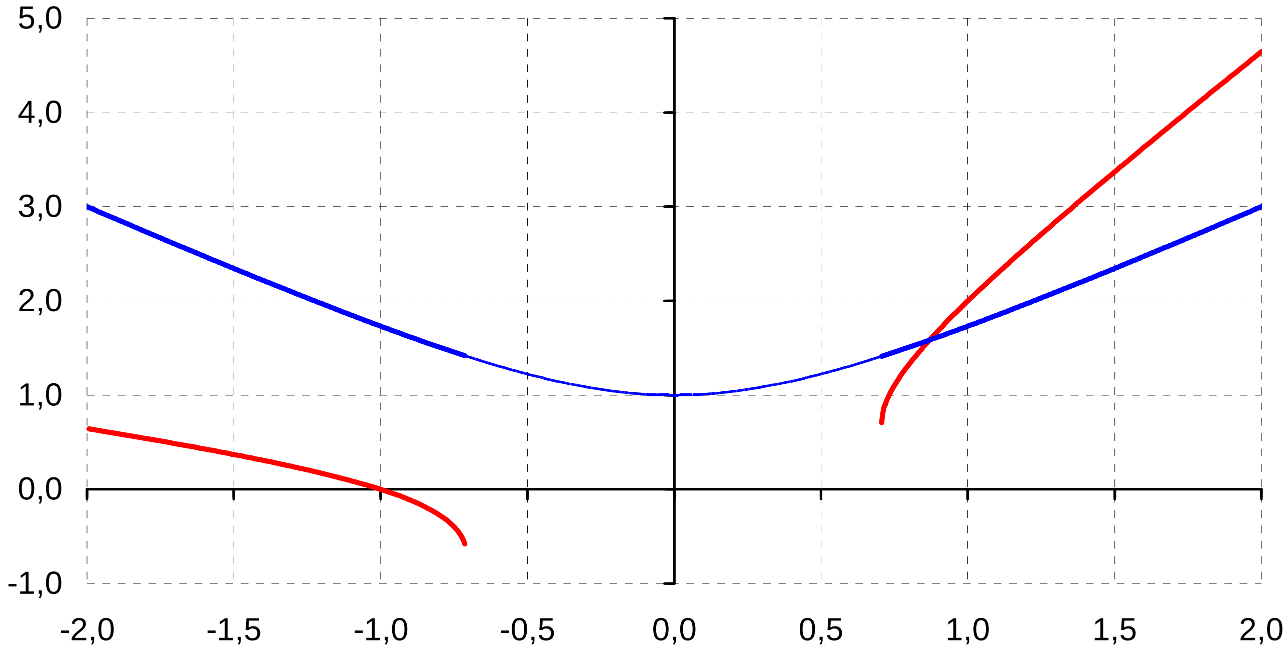
При приведении алгебраического уравнения (2.1) к канонической форме (2.2) будем иметь те же корни, что и для исходного уравнения.

Однако при этом могут появиться некоторые дополнительные корни.

Например, алгебраическое уравнение

$$\sqrt{2x^2-1}+x=\sqrt{2x^2+1}\,, \tag{2.3}$$

с областью определения $x \in (-\infty,-\sqrt{0.5}] \cup [\sqrt{0.5},\infty)$, имеет единственный корень – точка пересечения графиков функций $f_1(x)=\sqrt{2x^2-1}+x$ (график красного цвета с областью разрыва) и $f_2(x)=\sqrt{2x^2+1}$ (непрерывный график синего цвета):



Исходное уравнение (2.3), приведенное путем алгебраических преобразований к канонической форме:

$$7x^4-4=0 \quad \text{или} \quad x^4-\lambda^4=0, \quad \text{где } \lambda=\sqrt[4]{4/7}\,,$$

уже будет иметь дополнительные корни, включая и ***комплексно сопряженные***:

$$x^4-\lambda^4=(x^2-\lambda^2)(x^2+\lambda^2)=0\,,$$

откуда $$x_{1,2}=\pm\lambda \text{ и } x_{3,4}=\pm i\cdot\lambda\,.$$

Очевидно, что искомым решением исходного нелинейного уравнения (2.3) будет единственная точка с абсциссой $x_1=+\sqrt[4]{4/7}\approx 0.87$.

Если функция $F(x)$ не является алгебраической, то нелинейное уравнение (2.1) называется ***трансцендентным***.

Например, к трансцендентным уравнениям могут быть отнесены следующие нелинейные уравнения: $x - 10\sin(x) = 0$, $2^x - 2\cos(x) = 0$, $\lg(x+5) = \cos(x)$.

В некоторых частных случаях решение трансцендентных уравнений можно свести к решению им эквивалентных алгебраических уравнений.

Поскольку подавляющее большинство нелинейных уравнений с одной переменной не решаются путем аналитических преобразований (точными методами), на практике их решают только численным образом.

Решить такое уравнение – означает установить: имеет ли оно корни вообще, сколько этих корней, а затем – найти значения корней с требуемой точностью.

Задача численного определения действительных и комплексных корней уравнения (2.1) обычно состоит из двух этапов.

Первый этап – отделение корней, т.е. нахождение достаточно малых окрестностей рассматриваемой области, в которых гарантировано содержится один корень.

Второй этап – уточнение корней, т.е. вычисление абсциссы корня с заданной степенью точности.

В дальнейшем будем рассматривать численные методы нахождения действительных корней уравнения (2.1).

Наиболее распространенными на практике численными методами решения уравнения (2.1) являются: метод половинного деления, метод хорд, метод касательных (метод Ньютона), комбинированный метод, метод простой итерации и т.д.

Выбор того или иного численного метода для решения уравнения (2.1) зависит от искомого количества корней, способа задания исходного приближения к этим корням и от поведения функции $F(x)$ на общем интервале поиска корней.

Кроме того, если функция $F(x)$ задается таблично, то данная задача решается с привлечением интерполяционных многочленов, аппроксимирующих функцию $F(x)$.

Это, в свою очередь, обязывает более взвешенно подходить к вопросу выбора конкретного численного метода решения нелинейного уравнения и к проблеме программной реализации такого алгоритма в целом.

В этом случае итоговая погрешность при определении корня будет зависеть и от порядка точности, обеспечиваемой выбранным интерполяционным многочленом.

Проблема поиска корней комплексной функции может быть заменена процедурой поиска точки на комплексной плоскости, в которой достигает минимума модуль комплексной функции.

§ 2.1. Отделение корней.

Первый этап численного решения уравнения (2.1) состоит в отделении корней, т.е. в установлении достаточно «малых» промежутков из области определения функции $F(x)$, на которых гарантированно содержится только один простой корень.

Отделение корней во многих случаях можно произвести либо графически, либо воспользоваться вычисленной таблицей функции.

Принимая во внимание, что действительные корни уравнения (2.1) – это точки пересечения графика функции $F(x)$ с осью абсцисс, достаточно построить график функции $F(x)$ и отметить на оси Ox отрезки, содержащие по одному корню.

Построение графиков можно существенно упростить, заменив уравнение (2.1) равносильным ему уравнением $f_1(x) = f_2(x)$, как это мы выполнили выше для нелинейного уравнения (2.3).

В этом случае строятся графики функций $f_1(x)$ и $f_2(x)$, а потом на оси Ox отмечаются отрезки, локализирующие абсциссы точек пересечения этих графиков.

В сомнительных случаях графическое отделение корней необходимо подкреплять вычислениями.

При этом полезно использовать следующие очевидные положения:

- если непрерывная на отрезке $[a, b]$ функция $F(x)$ принимает на его концах значения разных знаков, т.е. $F(a) \cdot F(b) < 0$, то уравнение (2.1) имеет на этом отрезке, по меньшей мере, один корень;
- если функция $F(x)$ к тому же еще и строго монотонна (знак первой производной $F'(x)$ не меняется во всех точках на $[a, b]$), то корень уравнения (2.1) на отрезке $[a, b]$ будет единственным (простой корень).

Итак, пусть имеется уравнение $F(x) = 0$, причем можно считать, что все интересующие вычислителя корни находятся на некотором заданном отрезке $[A, B]$, на котором функция $F(x)$ определена и непрерывна.

Требуется отделить все корни уравнения $F(x) = 0$, т.е. указать все «малые» отрезки $[a_i, b_i] \subset [A, B]$, содержащие только по одному корню.

Будем вычислять значения $F(x)$ и $F'(x)$ (если это возможно), начиная с точки $x = A$, двигаясь вправо с некоторым шагом h.

Как только обнаружится пара соседних значений $F(x)$, имеющих разные знаки, и функция $F(x)$ будет монотонна на этом отрезке, так соответствующие значения аргумента x (предыдущее и последующее) можно считать концами отрезка $[a_i, b_i]$, содержащего корень x_i.

Очевидно, что надежность рассмотренного подхода к отделению корней уравнения зависит как от характера функции $F(x)$, так и от выбранной величины шага h продвижения от абсциссы A к абсциссе B.

Действительно, если при достаточно малом значении h на концах текущего отрезка $[x, x+h]$, функция $F(x)$ и ее производная $F'(x)$ не изменяют своих знаков, естественно ожидать, что уравнение $F(x)=0$ корней на этом отрезке не имеет.

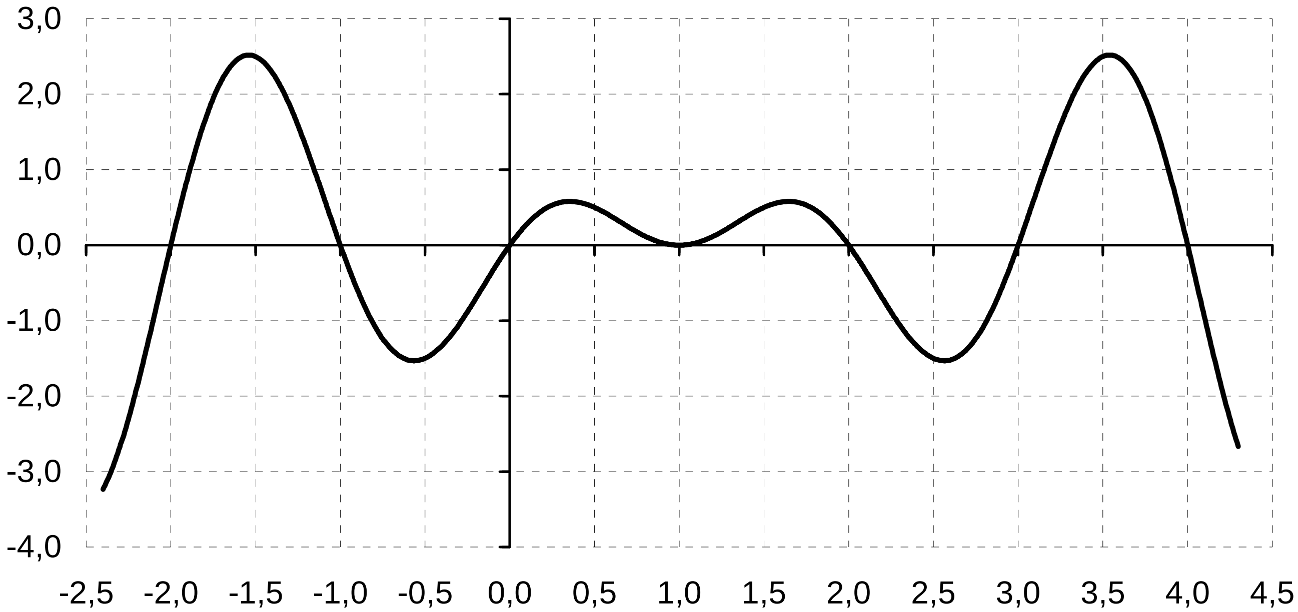
Если же условия монотонности функции $F(x)$ на отрезке $[x, x+h]$ не соблюдаются, т.е. $F'(x)\cdot F'(x+h)<0$, а необходимое условие $F(x)\cdot F(x+h)<0$ не выполнено, то возможно существование двух (или иное четное количество) корней на этом отрезке или имеется корень четной кратности, что проверяется отдельно.

Данная ситуация представлена ниже на графике функции

$$f(x)=(1-x)\sin(\pi x).$$

Так на отрезке $[\,0.5, 1.5\,]$ имеем $F(0.5)\cdot F(1.5)>0$ и $F'(0.5)\cdot F'(1.5)<0$.

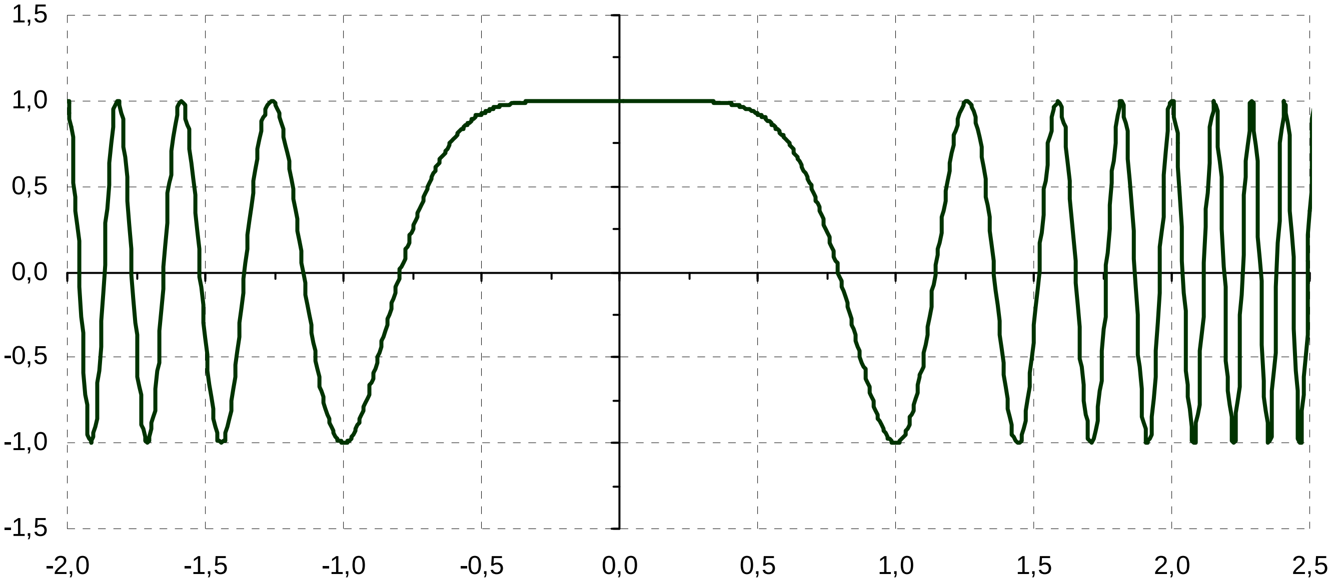
Визуальный анализ показывает, что указанная функция $f(x)$ имеет в точке $x_* = 1$ нуль кратности 2. При этом сам график функции ***касается*** оси абсцисс Ox в точке с координатой $x_* = 1$.



Остальные нули этой функции, представляющие собой множество целых чисел, очевидно, будут уже простыми корнями данного уравнения $(1-x)\sin(\pi x)=0$.

При одновременном выполнении условий $F(x)\cdot F(x+h)<0$ и $F'(x)\cdot F'(x+h)>0$ на отрезке $[x, x+h]$, последний может содержать как один корень, так и большее (нечетное) их количество.

Такое поведение характерно для сильно осциллирующих функций, например, для функции $f(x) = \cos(\pi x^3)$:



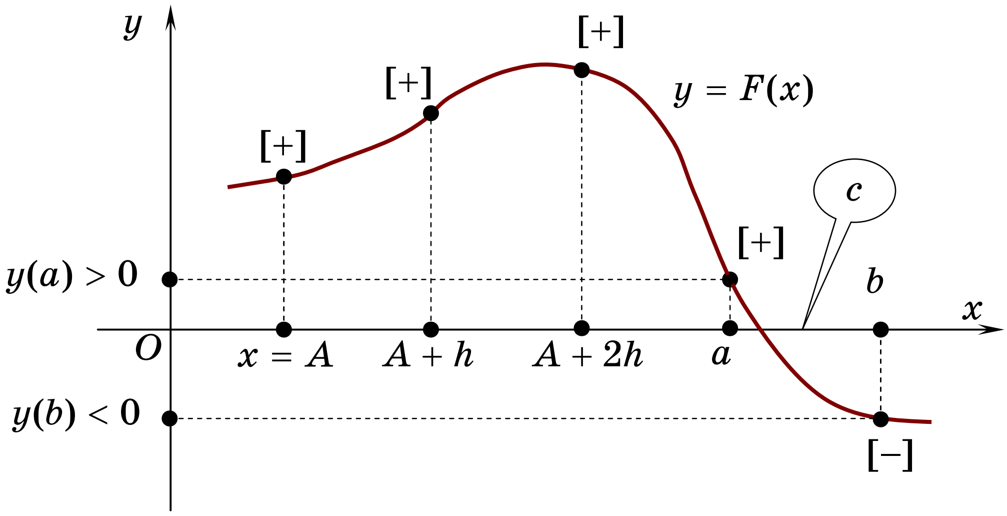
Предвидя подобные случаи, следует выбирать при отделении корней достаточно малые значения шага h, или проводить повторные уточняющие расчеты с шагом $h_1 = h / 10$ на выделенных (сомнительных) промежутках.

§ 2.2. Метод половинного деления (дихотомии).

Основная идея (алгоритм) метода дихотомии следующая.

Пусть на отрезке $[a, b]$ имеется единственный корень уравнения $F(x) = 0$.

Разделим этот отрезок пополам $[a, c] \cup [c, b]$ точкой $c = (a + b) / 2$.



Если $F(c) \neq 0$ (что практически наиболее вероятно), то возможны два случая: либо $F(x)$ меняет знак на первом отрезке $[a, c] \in [a, b]$, либо смена знака $F(x)$ происходит на втором отрезке $[c, b] \in [a, b]$.

Выбирая в каждом случае тот из отрезков, на котором функция меняет знак, и, продолжая процесс половинного деления дальше, приходим к сколь угодно малому отрезку, содержащему корень уравнения $F(x) = 0$.

Следовательно, итерационный процесс метода дихотомии можно представить в виде следующего цикла вычислений, где k – порядковый номер итерации.

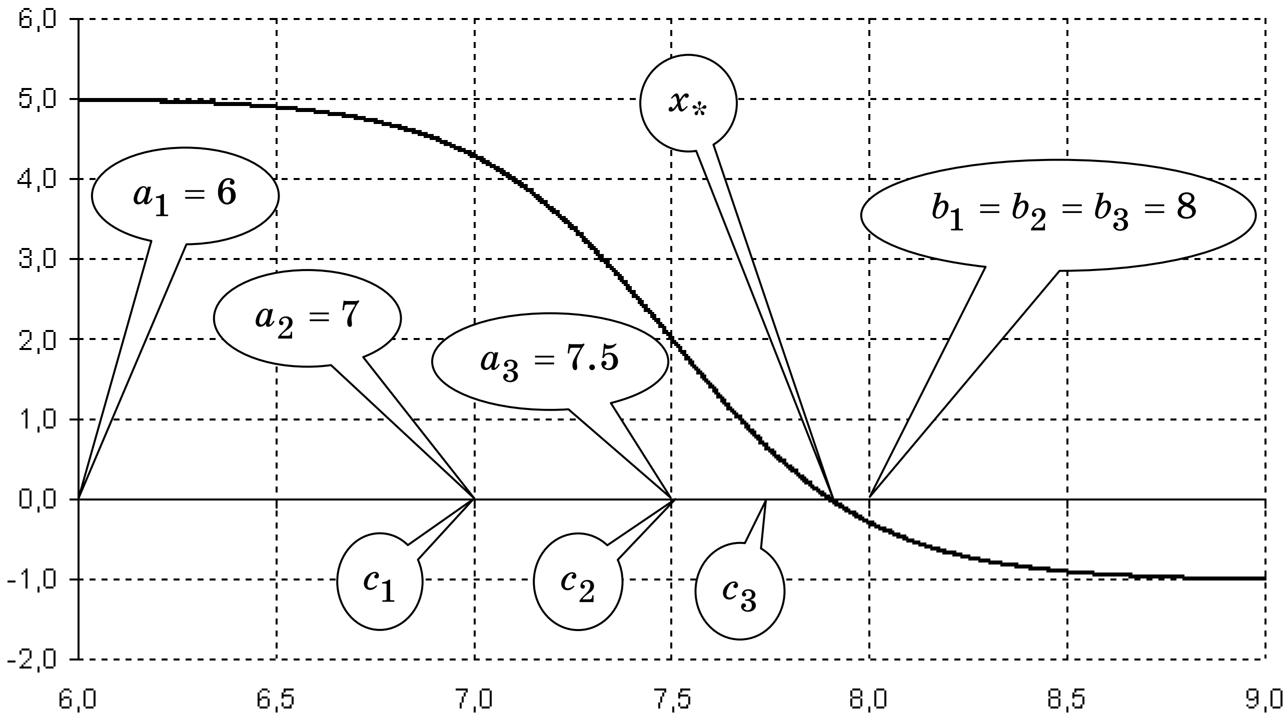
Пусть уже выделен отрезок $[a_k, b_k]$, на концах которого функция $F(x)$ имеет противоположные знаки и, следовательно, который содержит искомый корень x_*.

Определяем середину $c_k = (a_k + b_k)/2$ отрезка $[a_k, b_k]$ по алгоритму (1.4) и вычисляем значение $F(c_k)$.

При невыполнении условия $|F(c_k)| < \varepsilon$ итерации продолжаются по правилу:

$$\text{если } F(a_k)\cdot F(c_k) > 0, \begin{cases} \text{то} & a_{k+1} = c_k \text{ и } b_{k+1} = b_k, \\ \text{иначе} & a_{k+1} = a_k \text{ и } b_{k+1} = c_k, \end{cases}$$

т.е. строится новый отрезок $[a_{k+1}, b_{k+1}]$ с искомым корнем x_*.



На рисунке представлен итерационный процесс (первые три итерации) для уравнения $2 - 3\,\mathrm{th}(2x - 15) = 0$, когда в качестве начального отрезка, содержащего простой корень был выбран $[a_1, b_1] = [\ 6,\ 8\ ]$.

Поскольку искомый корень x_* здесь расположен близко к значению $b_1 = 8$, за первые три итерации «меняла» положение только левая граница интервала – a_k, а правая граница b_k – оставалась неизменной.

Данный метод является типично компьютерным, надежным и простым в реализации, хотя и обладает при этом медленной сходимостью.

Количество итераций K, необходимых для достижения заданной точности ε, можно оценить из соотношения $K > \log_2((b-a)/\varepsilon)$.

Алгоритм деления отрезка пополам, базирующийся на формуле (1.3-б), следует применять в программах, которые могут быть потенциально использованы для вычисления корней с максимально возможной точностью.

Кроме того, на процессорах, в которых вместо округления осуществляется отбрасывание лишних разрядов, формула (1.3-а) может вообще вывести за пределы текущего подотрезка.

Лучше придерживаться следующего правила, согласно которому новое приближение следует вычислять прибавлением полученной поправки к текущему приближению.

Положительная величина ε, задаваемая изначально, не должна быть меньше абсолютной погрешности вычисления самой функции $F(x)$ на интервале $[a,b]$.

Кроме этого, задаваемая точность ε должна выбираться в соответствии с количеством значащих цифр мантиссы чисел с плавающей запятой, использованных и соответствующим образом описанных в конкретной программе.

Если же на некотором шаге итераций имеет место $(b_k - a_k) < \varepsilon$, а условие $|F(c_k)| < \varepsilon$ не выполняется, то возможно зацикливание программы и расчеты необходимо принудительно завершать.

При этом итоговая погрешность вычислений, допущенная при определении корня, будет иметь величину $\varepsilon_1 = |F(c_k)|$, причем $\varepsilon_1 > \varepsilon$.

Выход из указанной ситуации возможен лишь путем алгебраического преобразования уравнения $F(x) = 0$ к какому-либо эквивалентному уравнению $G(x) = 0$ с последующим построением алгоритма повышенной точности для вычисления значений функции $G(x)$.

Метод дихотомии удобен при решении уравнений, для которых неизвестен аналитический вид функции $F(x)$.

Например, поиск отдельного собственного значения λ действительной матрицы $\mathrm{A}[n \times n]$ можно производить метом половинного деления, минуя стадию составления характеристического уравнения $f(\lambda) = 0$.

Для этого (при помощи соответствующих численных процедур) решается задача $\det(\mathrm{A} - \lambda \mathrm{E}) = 0$, где E – единичная матрица.

§ 2.3. Метод Ньютона (касательных).

Рассматриваемый метод базируется на теореме Больцано-Коши:

Если $F(x) \in C_2[a,b]$ и $F(a)\cdot F(b) < 0$, т.е. $F(x)$ принимает на концах отрезка $[a,b]$ значения с противоположными знаками, а $F'(x)$ не меняет знака на $[a,b]$, то уравнение $F(x)=0$ имеет на $[a,b]$ единственное решение (корень) x_*.

Существование решения $x_* \in [a,b]$ следует из непрерывности $F(x)$ на $[a,b]$ и предположения $F(a)\cdot F(b) < 0$. Не единственность решения при условии $F(a)\cdot F(b) < 0$ повлекла бы изменение знака у $F'(x)$ на $[a,b]$.

Метод Ньютона, называемый также методом касательных, состоит в следующем.

Рассмотрим в точке $x^{(0)}$ касательную к кривой $f = F(x)$, задаваемую уравнением $y_1(x) = F(x^{(0)}) + (x - x^{(0)})\cdot F'(x^{(0)})$.

Положив $y_1 = 0$, определяем точку $x^{(1)}$ пересечения данной касательной $y_1(x)$ с осью абсцисс Ox:

$$x^{(1)} = x^{(0)} - \frac{F(x^{(0)})}{F'(x^{(0)})}. \tag{2.4}$$

Строим новую касательную $y_2(x) = F(x^{(1)}) + (x - x^{(1)})\cdot F'(x^{(1)})$ к графику функции $f = F(x)$ в только что найденной точке $x^{(1)}$.

Эта прямая также пересекает ось абсцисс Ox.

Решаем новое линейное уравнение $y_2(x) = 0$ и получаем точку пересечения второй касательной с осью Ox:

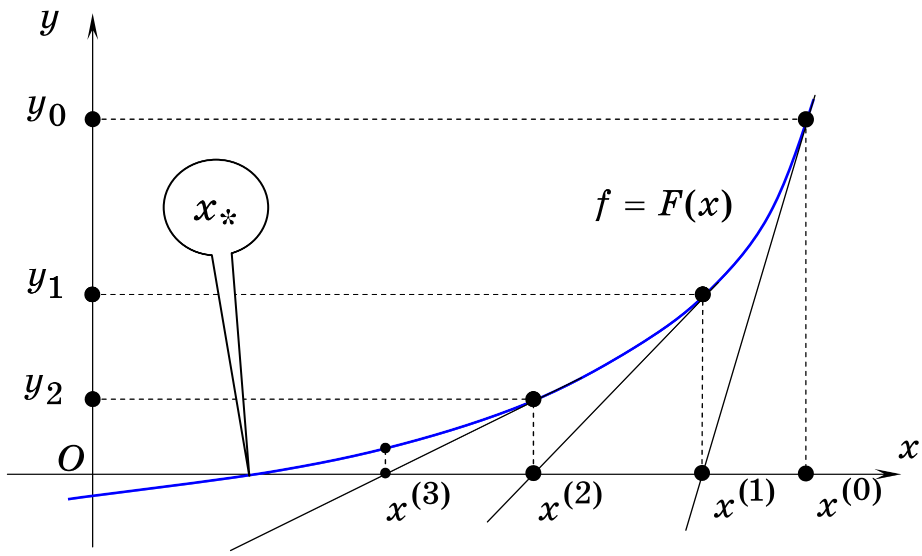
$$x^{(2)} = x^{(1)} - \frac{F(x^{(1)})}{F'(x^{(1)})}.$$

Очевидна итерационная последовательность данного численного метода:

$$y_k(x) = F(x^{(k)}) + (x - x^{(k)})\cdot F'(x^{(k)}), \qquad \Rightarrow$$

$$y_k(x^{(k+1)}) = 0 \quad \Rightarrow \quad x^{(k+1)} = x^{(k)} - \frac{F(x^{(k)})}{F'(x^{(k)})}. \tag{2.5}$$

Из геометрических соображений ясно, что при условиях сформулированной выше теоремы, итерационная последовательность $\{x^{(k)}\}$ по формуле (2.5), монотонно сходится к искомому решению x_* уравнения $F(x)=0$, см. рисунок на следующей странице.

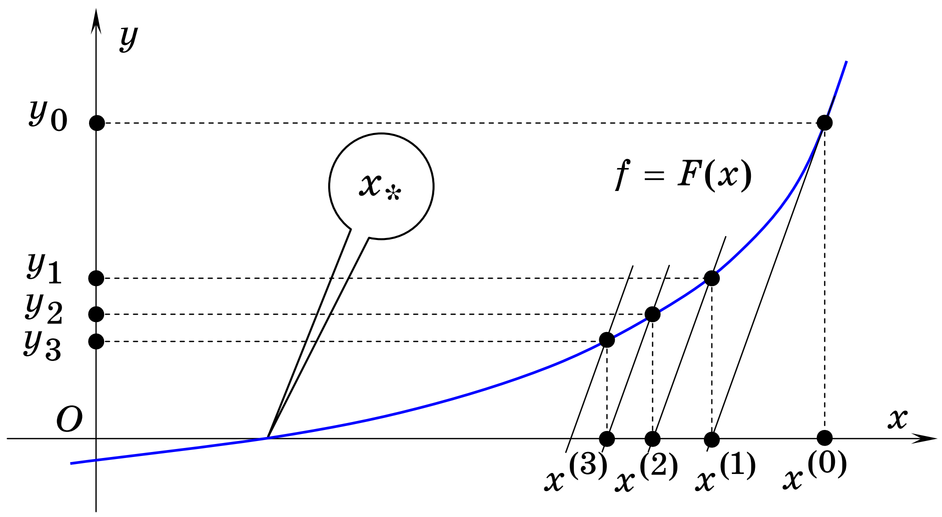


Итерации по схеме (2.5) для трансцендентных уравнений имеют наибольшую скорость сходимости среди известных численных методов.

Для простоты алгоритма иногда используют модифицированный метод Ньютона:

$$x^{(k+1)} = x^{(k)} - \frac{F(x^{(k)})}{F'(x^{(0)})}, \tag{2.6}$$

где упрощение состоит в том, что значение производной $F'(x^{(0)})$ (тангенс угла наклона касательной к графику функции) вычисляется один раз в начальной точке $x^{(0)}$.



Это, конечно, снижает скорость сходимости данного итерационного процесса, но никак не сказывается на его итоговой точности.

Очевидно, что пользоваться указанными итерационными методами (2.5)–(2.6) можно лишь в том случае, когда задан аналитический вид функции $F(x)$ и нет проблем с вычислением ее производных $F'(x)$ и $F''(x)$.

Определяющим моментом при реализации вычислений по данному методу является задание начального приближения $x^{(0)}$ к искомому корню x_*.

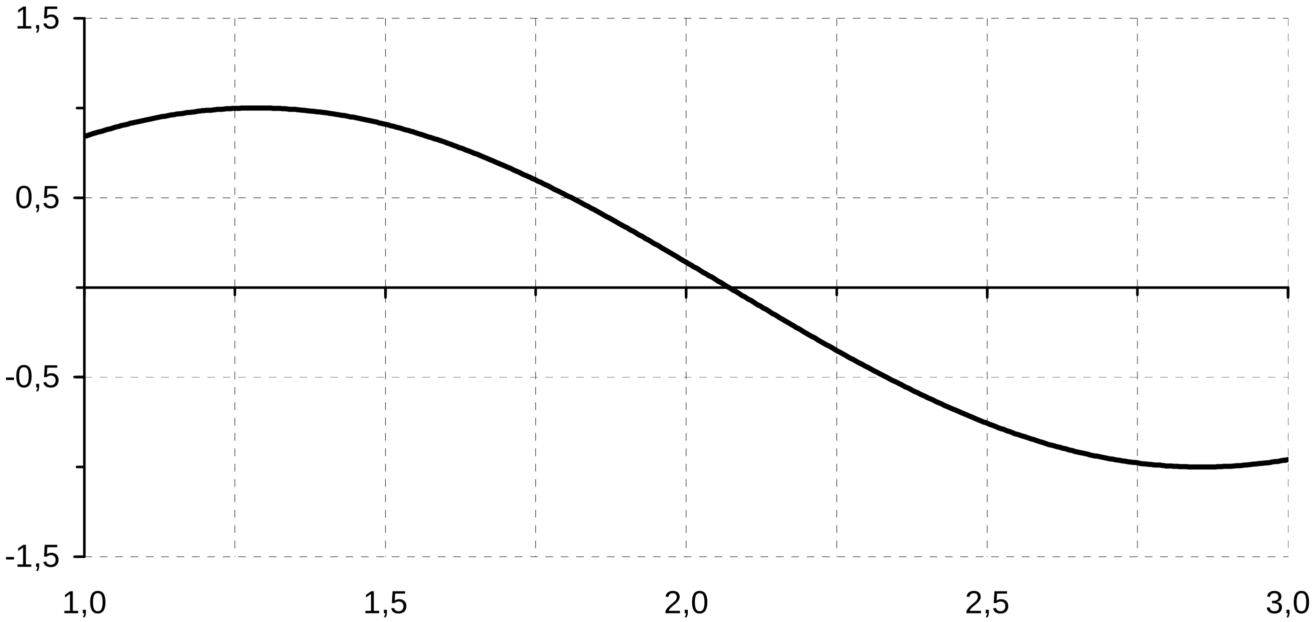
Поэтому, при выборе начального значения $x^{(0)}$ рекомендовано руководствоваться следующим дополнительным условием:

$$F(x^{(0)}) \cdot F''(x^{(0)}) > 0. \qquad (2.7)$$

Указанное условие не может быть использовано тогда, когда нуль функции $F(x)$ одновременно является нулем и для второй производной $F''(x)$.

В таком случае точка перегиба на графике функции совпадет с точкой пересечения графика функции и оси Ox.

Например, это будет для уравнения $F(x) = \sin(2x - 1) = 0$.



Очевидно, что в данном случае $F''(x) = -4\sin(2x-1) = -4 \cdot F(x)$, и тогда для любого значения переменной $x \in (-\infty, +\infty)$ будем иметь $F(x) \cdot F''(x) \le 0$.

Подводя итог сказанному, сформулируем основные положения метода касательных: если $F(a) \cdot F(b) < 0$, причем $F'(x)$ и $F''(x)$ отличны от нуля и сохраняют свои знаки на $[a, b]$, то при выборе начального приближения $x^{(0)}$ с учетом условия (2.7), итерационный процесс (2.5) всегда сходится к точному значению корня x_*.

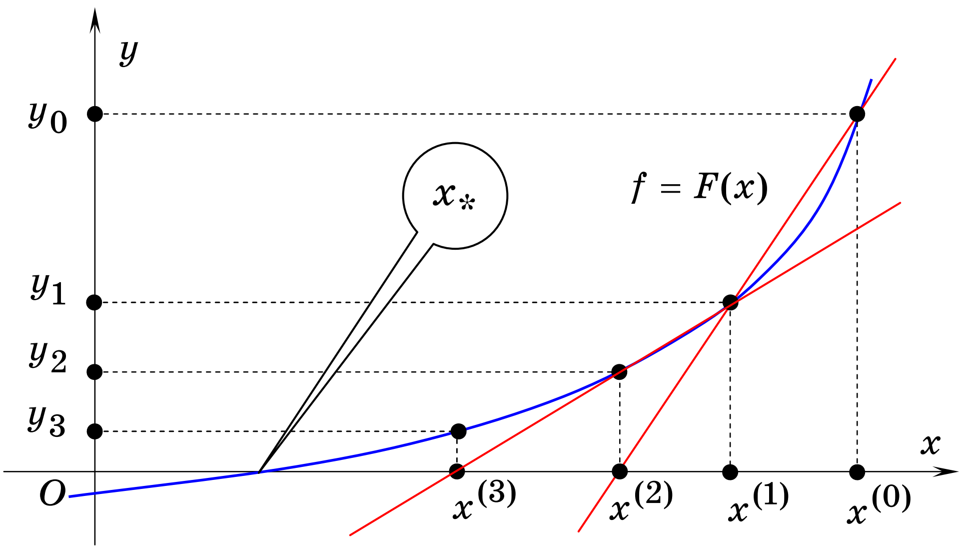
§ 2.4. Метод секущих (метод хорд).

Как уже отмечалось выше, в методе Ньютона необходимо вычислять производную $F'(x)$, что не всегда представляется возможным. Если заменить эту производную разделенной разностью первого порядка, вычисленной на двух последних итерациях:

$$F'(x^{(k)}) \approx \Delta F(x^{(k-1)}, x^{(k)}) = \frac{F(x^{(k)}) - F(x^{(k-1)})}{x^{(k)} - x^{(k-1)}}, \tag{2.8}$$

то из (2.5) можно получить итерационную формулу обобщенного метода хорд (секущих):

$$x^{(k+1)} = x^{(k)} - \frac{x^{(k)} - x^{(k-1)}}{F(x^{(k)}) - F(x^{(k-1)})} \cdot F(x^{(k)}), \quad k = 1, 2, \ldots \tag{2.9}$$



Геометрическая интерпретация классического метода хорд следующая: через две точки $(x^{(k-1)}, F(x^{(k-1)}))$ и $(x^{(k)}, F(x^{(k)}))$ кривой $f = F(x)$ проводится секущая, пересекающая ось абсцисс Ox в точке $x^{(k+1)}$.

Из узлов $x^{(k-1)}$ и $x^{(k)}$ отбрасывается тот, для которого значение функции $F(x)$ имеет тот же знак, что и значение $F(x^{(k+1)})$, или тот, для которого значение функции имеет большую величину (по модулю), после чего итерации (2.8) продолжаются.

В том случае, когда искомый корень x_* уже заключен в границах известного нам подотрезка $[a, b]$, поскольку для него производилась предварительная проверка на выполнение условия $F(a) \cdot F(b) < 0$, итерации по методу секущих можно упростить.

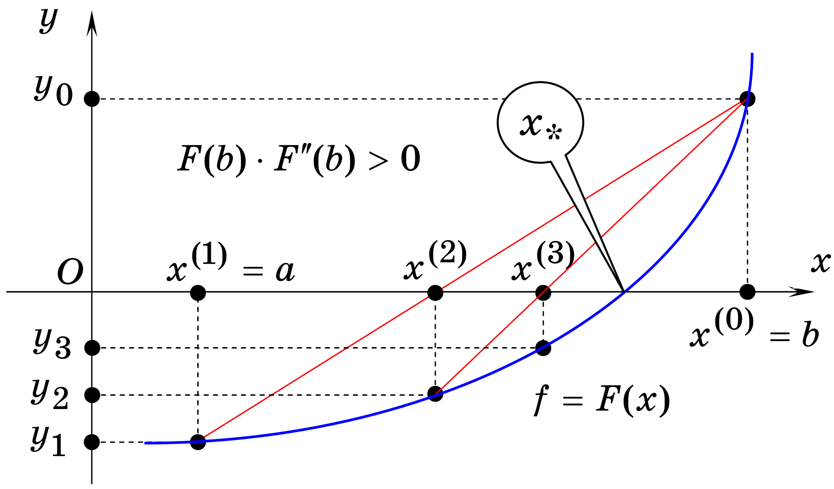
Очевидно, что если в качестве узлов $x^{(0)}$ и $x^{(1)}$ выбрать концы подотрезка $[a, b]$, т.е. взять, например, $x^{(0)} = a$ и $x^{(1)} = b$, то в процессе итераций (2.9) один из них не будет менять своего «положения».

Условимся, что в качестве $x^{(0)}$ мы будем выбирать тот из концов подотрезка $[a,b]$, для которого выполняется условие вида (2.7), т.е. $F(x^{(0)}) \cdot F''(x^{(0)}) > 0$.

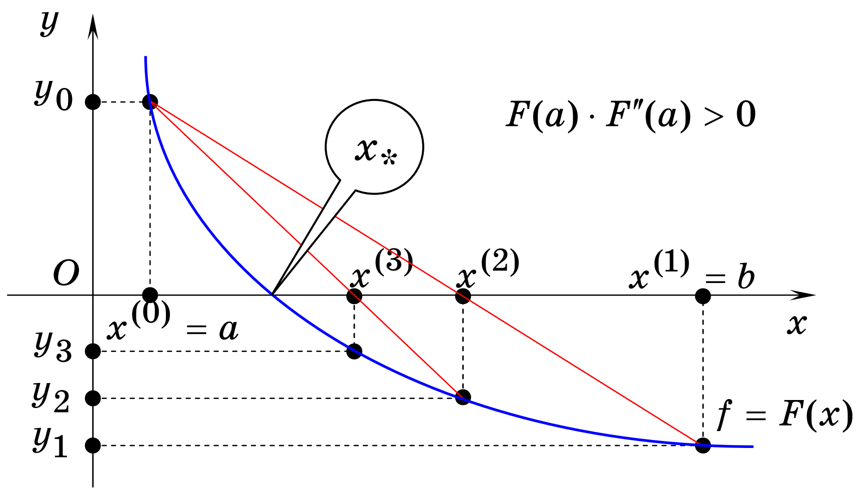
Таким образом, зафиксировав концевую точку $x^{(0)}$ подотрезка $[a,b]$, для которой выполняется условие $F(x^{(0)}) \cdot F''(x^{(0)}) > 0$, из (2.9) получим традиционную схему метода хорд (секущих):

$$x^{(k+1)} = x^{(k)} - \frac{x^{(k)} - x^{(0)}}{F(x^{(k)}) - F(x^{(0)})} \cdot F(x^{(k)}), \quad k = 1, 2, \ldots \tag{2.10}$$

Ниже представлена геометрическая интерпретация метода хорд по схеме (2.10), когда в качестве $x^{(0)}$ оказывается выбранной правая граница подотрезка $[a,b]$:



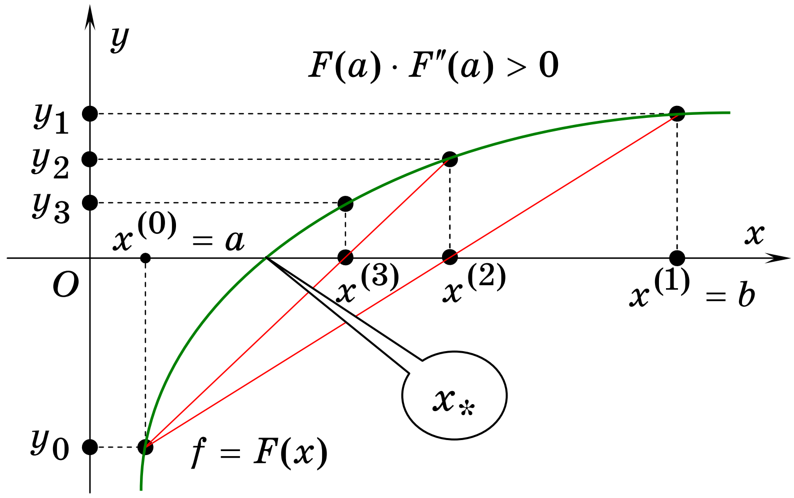
Альтернативная ситуация выглядит так:



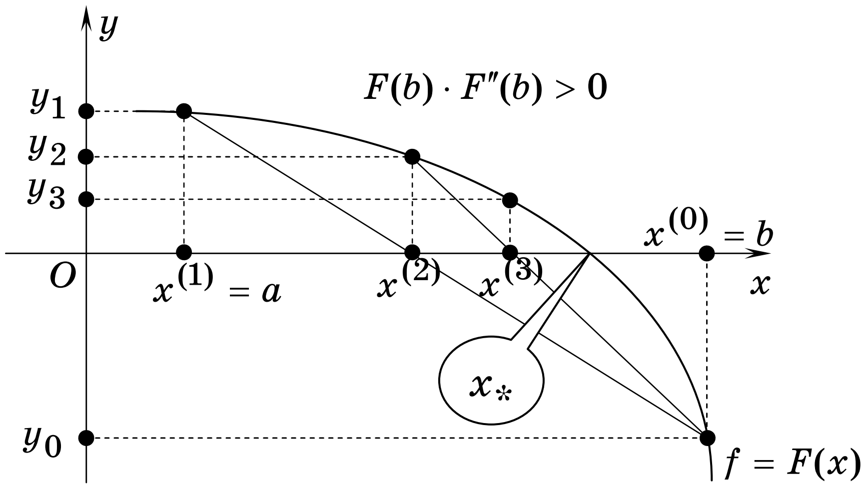
В рассмотренных выше двух случаях вторая производная $F''(x)$ функции $F(x)$ была положительной, т.е. $F''(x) > 0$ на $[a,b]$.

График функции $F(x)$ при этом на интервале $[a,b]$ имел вогнутость «***вниз***».

В тех случаях, когда $F''(x) < 0$ на интервале $[a,b]$ (т.е. график функции имеет вогнутость «***вверх***»), очевидно, возможны еще две геометрические интерпретации хода итераций по методу секущих:



или



Сложность с выбором $x^{(0)}$ может возникнуть, если на сегменте $[a,b]$ меняет знак вторая производная. Такое уравнение мы уже рассматривали выше – уравнение $\sin(2x-1) = 0$. В данном случае всегда имеет место $F(x) \cdot F''(x) \le 0$.

Или может оказаться, что всегда $F(x) \cdot F''(x) \ge 0$, например, для алгебраического (нелинейного) уравнения $F(x) = (x-1)^3 = 0$.

Метод касательных или метод секущих может не обеспечить сходимости для корня уравнения такого типа.

В таких случаях следует воспользоваться методом дихотомии для уточнения корня.

Для определения кратных нулей нелинейного уравнения можно воспользоваться каким-либо приближенным (численным) методом поиска точек минимума для функции $\Phi(x) = |\,F(x)\,|$. Очевидно, что если в некоторой точке $\tilde{x} \in [a,b]$ функция $F(\tilde{x}) = 0$, то функция $\Phi(x)$ имеет в этой точке минимум, притом равный нулю, т.е. $\Phi(\tilde{x}) = 0$.