Chapter I: Numerical Solution of Nonlinear Equations with a Single Variable

In this chapter, we will study three methods for solving nonlinear equations with one variable, also known as transcendental equations. Examples of these equations include:

$\boldsymbol{f(x) = sin(x) + x = 0}$, $\boldsymbol{f(x) = ln(x) - 2x + 3 = 0}$. These equations do not have exact roots that can be calculated directly, which is why we resort to numerical methods to find approximate solutions. The calculated roots are as precise as desired, especially when computational resources are available.

These numerical methods only allow the calculation of a single root on a well-chosen interval. Therefore, if the equation has more than one root, it is necessary to locate them in carefully chosen intervals and perform the calculation for each root separately.

1. Locating the Roots of an Equation f(x) = 0

Consider an equation f(x) = 0 for which we seek the solution on an interval [a, b]. We begin by making a rough sketch of the function on the given interval and then isolate each root in a subinterval that is as narrow as possible. Figure 1 shows the graph of a function f that intersects the x-axis at three points, which means that the equation f(x) = 0 has three roots. We denote the exact roots by $\boldsymbol{\bar{x}_1}, \boldsymbol{\bar{x}_2}$ and $\boldsymbol{\bar{x}_3}$.

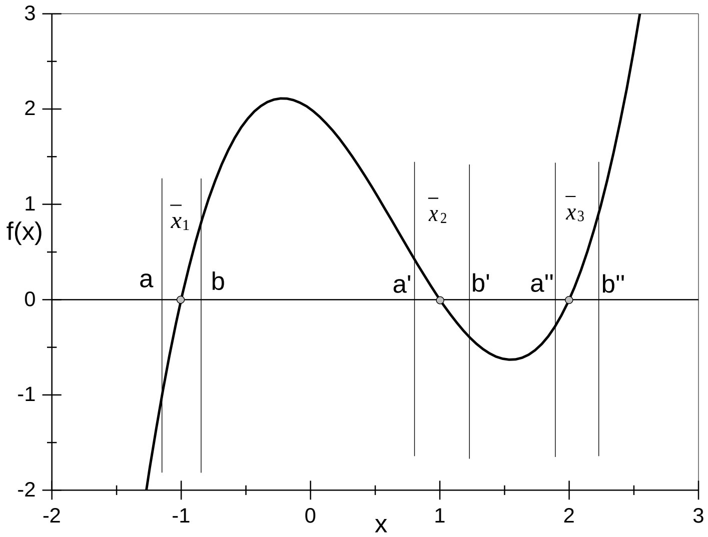


Figure 1.1: Illustration of root localization

We observe that the function is continuous on each subinterval, and each subinterval:

- Contains only one root such that $\boldsymbol{\bar{x}_1} \in [\boldsymbol{a} \;\; , \boldsymbol{b} \;\;], \boldsymbol{\bar{x}_2} \in [\boldsymbol{a'}, \boldsymbol{b'}]$et $\boldsymbol{\bar{x}_3} \in [\boldsymbol{a''}, \boldsymbol{b''}]$.
- Satisfies the condition $\boldsymbol{f(a)f(b) < 0}$, $\boldsymbol{f(a')f(b') < 0}$ **et** $\boldsymbol{f(a'')f(b'') < 0}$.

The form of the equation f(x) = 0 can be complicated. In this case, if possible, it can be decomposed into two simpler parts: ***g(x) = h(x).***

For example, the equation $\boldsymbol{f(x) = \ln(x) - x^2 + 2 = 0}$, which is quite complicated to plot, can be decomposed into: $\boldsymbol{g(x) = h(x)}$ with $\boldsymbol{g(x) = \ln(x)}$ **and** $\boldsymbol{h(x) = x^2 - 2}$

The graphs of g and h are very simple. The solution of ***f(x) = 0*** is located at the intersection of ***g*** and ***h***. Then, we project the intersection points onto the x-axis and locate the roots. We can easily verify the intervals found by calculating $\boldsymbol{f(a_i)f(b_i) < 0}$ for each interval [a_i, b_i].

<u>Example</u>:

Consider the equation: $\ln(x) - x^2 + 2 = 0$. Let's locate its roots. At this point, we do not know the number of roots of this equation.

We plot the curve of the function: $\ln(x) - x^2 + 2 = 0$. The intersections of the curve with the x-axis represent the roots of this function.

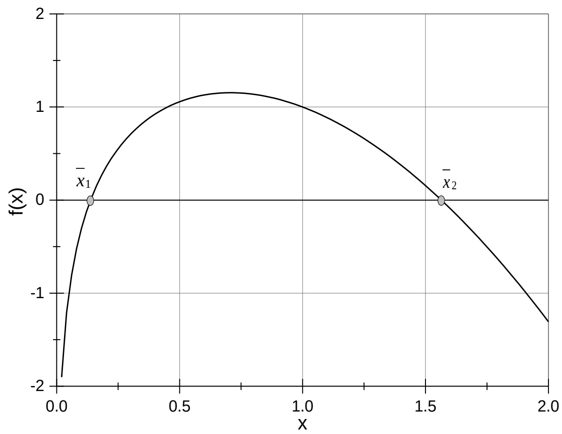


Figure 1.2: Graph of function f

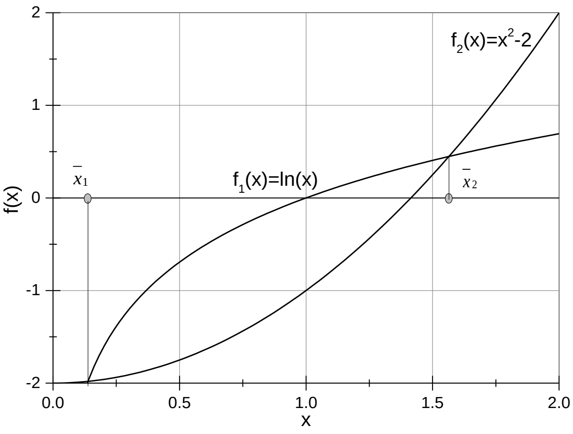


Figure 1.3: Graphs of functions f_1 and f_2

We note that this equation has two roots (Fig. 2) that belong, for example, to the intervals [0.1, 0.5] and [1, 2].

We can also rewrite the function ***f(x) = 0*** in a simpler form, for example:

$\ln(x) - x^2 + 2 = 0 \Leftrightarrow \ln(x) = x^2 - 2$

or $f_1(x) = f_2(x)$; with $f_1(x) = \ln(x)$ and $f_2(x) = x^2 - 2$

Then we plot these two functions (Fig. 3), which are easy to plot on the same axes. Their intersections represent the roots of f(x) = 0.

<u>2. Bisection or Dichotomy Method</u>

This is the simplest method and requires the most calculations. It is based on the special case (Bolzano's Theorem) of the intermediate value theorem, which states that:

1. If ***f(x)*** is continuous on the interval ***[a, b],***
2. If ***f(a)*** and ***f(b)*** do not have the same sign, then there exists at least one real number c between ***a*** and ***b*** such that ***f(c) = 0*** (since ***0*** is between ***f(a)*** and ***f(b)***).

2.1 Principle of the Method

Once the roots are located, each in an interval, for the sake of simplicity of writing, let's take for example [a, b]:

1. We divide the interval into two equal parts such that $x_0 = \frac{a+b}{2}$.
2. We obtain two subintervals $[a, x_0]$ and $[x_0, b]$. The root $\bar{x}$ must necessarily belong to one of them. To verify, we calculate $f(a)f(x_0)$ and $f(x_0)f(b)$,. The negative product is the one that corresponds to the interval that contains the solution.

3. We denote the new interval by $[\boldsymbol{a_1}, \boldsymbol{b_1}]$ (Fig. 4) such that:

$$\boldsymbol{a_1} = \begin{cases} \boldsymbol{a}\ \boldsymbol{si}\ \bar{\boldsymbol{x}} \in [\boldsymbol{a}, \boldsymbol{x_0}] \\ \boldsymbol{x_0}\ \boldsymbol{si}\ \bar{\boldsymbol{x}} \in [\boldsymbol{x_0}, \boldsymbol{b}] \end{cases} \qquad et \qquad \boldsymbol{b_1} = \begin{cases} \boldsymbol{x_0}\ \boldsymbol{si}\ \bar{\boldsymbol{x}} \in [\boldsymbol{a}, \boldsymbol{x_0}] \\ \boldsymbol{b}\ \boldsymbol{si}\ \bar{\boldsymbol{x}} \in [\boldsymbol{x_0}, \boldsymbol{b}] \end{cases}$$

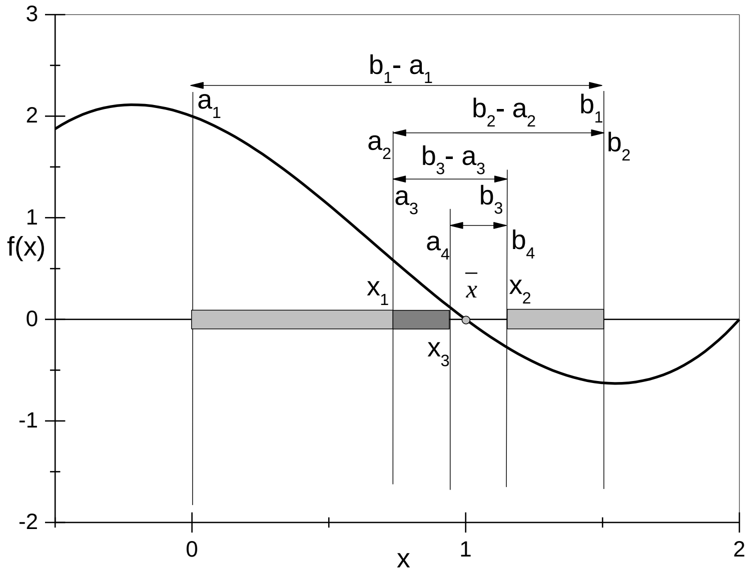


Figure 1.4: Illustration of the bisection method

By repeating (iterating) the same method for the interval obtained, we will have the values:

$$\boldsymbol{x_1} = \frac{\boldsymbol{a_1} + \boldsymbol{b_1}}{\boldsymbol{2}}, \boldsymbol{x_2} = \frac{\boldsymbol{a_2} + \boldsymbol{b_2}}{\boldsymbol{2}}, \dots\dots\dots\dots, \qquad \boldsymbol{x_n} = \frac{\boldsymbol{a_n} + \boldsymbol{b_n}}{\boldsymbol{2}}$$

The sequence $\{\boldsymbol{x_n}\}_{\boldsymbol{n=0},\infty}$ converges to the solution $\bar{\boldsymbol{x}}$ of f(x) = 0 when n → ∞.

2.2 Number of Divisions to Achieve a Given Precision ε

Since each time we divide the interval into two equal parts, we have:

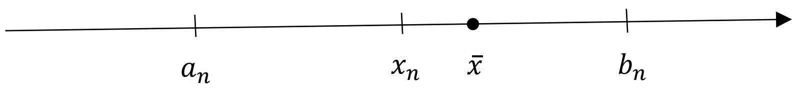
$$\boldsymbol{b_1} - \boldsymbol{a_1} = \frac{\boldsymbol{b} - \boldsymbol{a}}{\boldsymbol{2}};$$

$$\boldsymbol{b_2} - \boldsymbol{a_2} = \frac{\boldsymbol{b_1} - \boldsymbol{a_1}}{\boldsymbol{2}} = \frac{\boldsymbol{1}}{\boldsymbol{2}}\frac{\boldsymbol{b} - \boldsymbol{a}}{\boldsymbol{2}} = \frac{\boldsymbol{b} - \boldsymbol{a}}{\boldsymbol{2^2}};$$

$$\boldsymbol{b_3} - \boldsymbol{a_3} = \frac{\boldsymbol{b_2} - \boldsymbol{a_2}}{\boldsymbol{2}} = \frac{\boldsymbol{1}}{\boldsymbol{2}}\frac{\boldsymbol{b} - \boldsymbol{a}}{\boldsymbol{2^2}} = \frac{\boldsymbol{b} - \boldsymbol{a}}{\boldsymbol{2^3}};$$

....

$$\boldsymbol{b_n} - \boldsymbol{a_n} = \frac{\boldsymbol{b_{n-1}} - \boldsymbol{a_{n-1}}}{\boldsymbol{2}} = \frac{\boldsymbol{1}}{\boldsymbol{2}}\frac{\boldsymbol{b} - \boldsymbol{a}}{\boldsymbol{2^{n-1}}} = \frac{\boldsymbol{b} - \boldsymbol{a}}{\boldsymbol{2^n}}.$$



Puisque $\bar{\boldsymbol{x}} \in [\boldsymbol{a_n}, \boldsymbol{b_n}] = [\boldsymbol{a_n}, \boldsymbol{x_n}] \cup [\boldsymbol{x_n}, \boldsymbol{b_n}]$ on a $|\boldsymbol{x_n} - \bar{\boldsymbol{x}}| \leq \frac{\boldsymbol{1}}{\boldsymbol{2}}\frac{\boldsymbol{b-a}}{\boldsymbol{2^n}} = \frac{\boldsymbol{b-a}}{\boldsymbol{2^{n+1}}}$

The difference $|\boldsymbol{x_n} - \overline{\boldsymbol{x}}|$, which is the error of the calculation, must be less than a given precision ε, that is:

$$|\boldsymbol{x_n} - \overline{\boldsymbol{x}}| \leq \boldsymbol{\varepsilon}$$

Then, it suffices that $\frac{\boldsymbol{b-a}}{\boldsymbol{2^{n+1}}} \leq \boldsymbol{\varepsilon}$

This gives $n \geq \frac{\ln\left(\frac{b-a}{2\varepsilon}\right)}{\ln(2)}$

- The number of divisions depends only on the length of the interval and the precision.
- This method is unconditionally convergent. Its problem is that it is slow, which is why it is used to start other more elaborate methods.

Example:

Let's calculate the first root of the equation $\mathbf{ln}(\boldsymbol{x}) - \boldsymbol{x^2} + \mathbf{2} = \mathbf{0}$, which belongs to [0.1, 0.5], with a precision of 0.01.

Let's calculate the number of divisions to make:

$$\boldsymbol{n} \geq \frac{\ln\left(\frac{b-a}{2\varepsilon}\right)}{\ln(2)} = \frac{\ln\left(\frac{0.5-0.1}{2*0.01}\right)}{\ln(2)} = \mathbf{4.32}$$

We take n = 5 since n is an integer and greater than 4.32.

$$f(a_1) = f(0.1) = -0.313 \quad \text{et} \quad f(b_1) = f(0.5) = 1.057$$

$$x_1 = \frac{a_1 + b_1}{2} = \frac{0.1 + 0.5}{2} = 0.30; \quad f(0.3) = 0.706 > 0 \text{ thus } a_2 = 0.1 \text{ et } b_2 = 0.3$$

$$x_2 = \frac{a_2 + b_2}{2} = \frac{0.1 + 0.3}{2} = 0.20; \quad f(0.2) = 0.351 > 0 \text{ thus } a_3 = 0.1 \text{ et } b_3 = 0.2$$

$$x_3 = \frac{a_3 + b_3}{2} = \frac{0.1 + 0.2}{2} = 0.15; \quad f(0.15) = 0.080 > 0 \text{ thus } a_4 = 0.1 \text{ et } b_4 = 0.15$$

$$x_4 = \frac{a_4 + b_4}{2} = \frac{0.1 + 0.15}{2} = 0.125; \quad f(0.125) = -0.095 < 0 \text{ thus } a_5 = 0.15 \text{ et } b_4 = 0.125$$

$$x_5 = \frac{a_5 + b_5}{2} = \frac{0.15 + 0.125}{2} = 0.1375; \quad f(0.125) = -0.003 \text{ the solution is } \overline{x} \cong x_5 = 0.1375$$

3. Method of Successive Approximations or Fixed Point Method

Let $\boldsymbol{g}$ be a function defined on an interval ***[a, b]***. The point $\overline{\boldsymbol{x}}$ which satisfies $\overline{\boldsymbol{x}}$ = g($\overline{\boldsymbol{x}}$) with $\overline{\boldsymbol{x}} \in$ [a, b] is called a fixed point of the function g.

This method is based on the principle of the fixed point of a function. We write the equation ***f(x) = 0*** in the form ***x = g(x)***, then we look for the fixed point $\overline{\boldsymbol{x}}$ of the function g. For this, we create the sequence $x_{n+1} = g(x_n)$ (n=0,1,2….) with x_0 given by dichotomy for example.

We start from x_0 for *n = 0*, we calculate $\boldsymbol{x_1} = \boldsymbol{g}(\boldsymbol{x_0})$, then for *n = 1*, we calculate $\boldsymbol{x_2} = \boldsymbol{g}(\boldsymbol{x_1})$,..., $\boldsymbol{x_{n+1}} = \boldsymbol{g}(\boldsymbol{x_n})$. Under certain conditions, the sequence $\{\boldsymbol{x_n}\}_{\boldsymbol{n=0},\infty}$ converges towards the solution $\overline{\boldsymbol{x}}$, fixed point of $\boldsymbol{g}$ and solution of the equation *f(x)=0*.

Example:

Write the equation ***f(x) = 0*** in the form ***x = g(x)*** if $\boldsymbol{f(x) = x^2 + 3e^x - 12}$.

We can write:

- $\boldsymbol{x = g_1(x) = x^2 + 3e^x - 12 + x}$
- $\boldsymbol{x = g_2(x) = \sqrt{12 - 3e^x}}$
- $\boldsymbol{x = g_3(x) = ln\left(\frac{12-x^2}{3}\right)}$

To be able to choose the appropriate form of $\boldsymbol{g}$ for the calculation, a convergence criterion of this method must be verified.

3.1 Convergence Criterion and Calculation Stop Criterion for the Method of Successive Approximations

Let g be a derivable function defined from ***[a, b]*** to ***[a, b]*** such that (sufficient condition):

$$\boldsymbol{|g'(x)| \leq k < 1 \quad \forall\, x \in [a, b]}$$

Then the sequence $\{\boldsymbol{x_n}\}_{\boldsymbol{n=0},\infty}$ defined by $\boldsymbol{x_{n+1} = g(x_n)}$ **(n=0,1,2….)** converges independently of the value of $\boldsymbol{x_0}$ to the unique fixed point $\boldsymbol{\bar{x}}$ of $\boldsymbol{g}$.

If several forms of $\boldsymbol{g}$ satisfy this condition, we will have several values of $\boldsymbol{k}$. We choose the one with the minimum value of $\boldsymbol{k}$. In practice, we calculate $\boldsymbol{k = max_{x\in[a,b]}|g'(x)|}$ which must be less than one for the method to converge.

We stop the calculations for this method when the absolute difference between two successive iterations is less than a certain given precision ε.

$$\boldsymbol{|x_{n+1} - x_n| < \varepsilon}$$

Example:

Find the first root of the equation $\boldsymbol{\ln(x) - x^2 + 2 = 0}$ which belongs to **[0.1, 0.5]** with a precision **ε = 0.001.** We write this equation in the form ***x = g(x)*** and verify the convergence conditions. We can write:

$\boldsymbol{x = e^{x^2-2} = g_1(x)}$ and

$\boldsymbol{x = \sqrt{ln(x) + 2} = g_2(x)}$

Let's verify the convergence condition for this method: $\boldsymbol{k = max_{x\in[a,b]}|g'(x)|}$.

For $\boldsymbol{x = e^{x^2-2} = g_1(x)}$, we have $\boldsymbol{k_1 = max_{x\in[0.1,0.5]}|g_1{}'(x)| = max_{x\in[0.1,0.5]}\left|2xe^{x^2-2}\right|}$. On the interval [0.1, 0.5], $\boldsymbol{g_1{}'(x)}$ is strictly increasing, so the maximum value is $\boldsymbol{k_1 = max_{x=0.5}\left|2 * 0.5e^{0.5^2-2}\right| = 0.174 < 1}$. Thus, this form converges.

We write: $\boldsymbol{x_{n+1} = g_1(x_n) = e^{x_n^2-2}}$ **(n=0,1,2,…..)**

Let's start with $\mathbf{x_0 = 0.3}$, the midpoint of the given initial interval:

We calculate

$$\boldsymbol{n=0}, \quad \boldsymbol{x_1 = g_1(x_0) = e^{x_0^2-2}} = 0.148$$

We calculate $|x_1 - x_0| = 0.152 > \varepsilon$;

$$\boldsymbol{n=1}, \ \boldsymbol{x_2 = g_1(x_1) = e^{x_1^2-2}} = 0.138.$$

We calculate $|x_2 - x_1| = 0.01 > \varepsilon$

$$\boldsymbol{n=2}, \ \boldsymbol{x_3 = g_1(x_2) = e^{x_2^2-2}} = 0.138.$$

We calculate $|x_3 - x_2| = 0.00 < \varepsilon,$

The solution is $\boldsymbol{\bar{x}} \approx \boldsymbol{x_3} = 0.254$.

4. Newton-Raphson Method

This is the most efficient and most used method. It is based on the Taylor series expansion. If ***f(x)*** is continuous and continuously differentiable in the neighborhood of $\boldsymbol{\bar{x}}$, the solution of ***f(x) = 0***, then the Taylor series expansion around an estimate x_n close to $\boldsymbol{\bar{x}}$ is written:

$$f(\bar{x}) = f(x_n) + \frac{(\bar{x} - x_n)}{1!} f'(x_n) + \frac{(\bar{x} - x_n)^2}{2!} f''(x_n) + \cdots$$

If x_n is a close estimate of $\boldsymbol{\bar{x}}$, then the square of the error $\boldsymbol{\varepsilon_n = \bar{x} - x_n}$ and the terms of higher degrees are negligible. Knowing that $f(\bar{x}) = 0$, we obtain the approximate relation:

$$\boldsymbol{f(x_n) + (\bar{x} - x_n) f'(x_n) \approx 0}$$

Therefore, $\boldsymbol{\bar{x} = x_n - \frac{f(x_n)}{f'(x_n)}}$

We can write the $(\boldsymbol{n+1})^{\boldsymbol{th}}$ iteration approximating $\boldsymbol{\bar{x}}$ as:

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

This sequence, if it converges, must converge towards the solution $\boldsymbol{\bar{x}}$ of *f(x) = 0*. We note that ***f'(x)*** must be non-zero.

4.1 Convergence Criterion of the Newton-Raphson Method

Let *f* be a function defined on *[a, b]* such that:

i. $f(a)f(b) < 0$

ii. *f'(x)* and *f''(x)* are non-zero and maintain a constant sign on the given interval.

4.2 Calculation Stop Criterion for the Newton-Raphson Method

If the convergence condition is verified, the iterative process must converge. This means that each new iteration is better than the previous one. Therefore, we can say that if we have a precision ε, we stop the calculation when the absolute difference between two successive approximations is less than the given precision. That is:

$$\boldsymbol{|x_{n+1} - x_n| \leq \varepsilon}$$

If this condition is verified, we take $\boldsymbol{x_{n+1}}$ as the solution of *f(x) = 0*.

Example:

Find the first root of the equation $\boldsymbol{f(x) = \ln(x) - x^2 + 2 = 0}$ which belongs to ***[0.1, 0.5]*** with a precision ***ε = 0.0001***. We calculate the first and second derivatives of *f* and verify the convergence conditions.

We have: $\boldsymbol{f'^{(x)} = \frac{1}{x} - 2x > 0,}$ *f'(x)* is strictly decreasing and positive on the given interval.

$\boldsymbol{f''(x) = -\frac{1}{x^2} - 2, f''(x) < 0}$ on the given interval.

The convergence condition is verified. We therefore write:

$$\boldsymbol{x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{\ln(x_n) - x_n{}^2 + 2}{\frac{1}{x_n} - 2x_n}}, \text{ (n=0,1,2,…..).}$$

Let's start with $x_0 = 0.3$, the midpoint of the given initial interval:

$$\boldsymbol{n = 0},\ \boldsymbol{x_1 = x_0 - \frac{\ln(x_0) - x_0{}^2 + 2}{\frac{1}{x_0} - 2x_0} = 0.0417} \quad |x_1 - x_0| > \varepsilon\,;$$

$$\boldsymbol{n = 1},\ \boldsymbol{x_2} = 0.0910.\ |x_2 - x_1| > \varepsilon$$

$$\boldsymbol{n = 2},\ \boldsymbol{x_3} = 0.1285.\ |x_3 - x_2| > \varepsilon$$

$$\boldsymbol{n = 3},\ \boldsymbol{x_4} = 0.1376.\ |x_4 - x_3| > \varepsilon$$

$$\boldsymbol{n = 4},\ \boldsymbol{x_5} = 0.1379.\ |x_5 - x_4| > \varepsilon$$

$$\boldsymbol{n = 5},\ \boldsymbol{x_6 = 0.1379}. \text{ The solution is } \boldsymbol{x_6 = 0.1379}$$

Remarks:

- The bisection method is unconditionally convergent. Its disadvantage is its slowness to obtain the solution with high precision. It can be used to start other more efficient methods.
- The method of successive approximations is faster than the bisection method provided that it converges.
- The Newton-Raphson method is the fastest. It allows obtaining very precise solutions in a reduced number of iterations.