"Socrates dialectical Procedure: For an over all view what is now necessary is the movement of consciousness from knowledge of particular objects to an understanding of general concepts."

Socrates (469-399) BCE

Chapter 2

Equation Solving

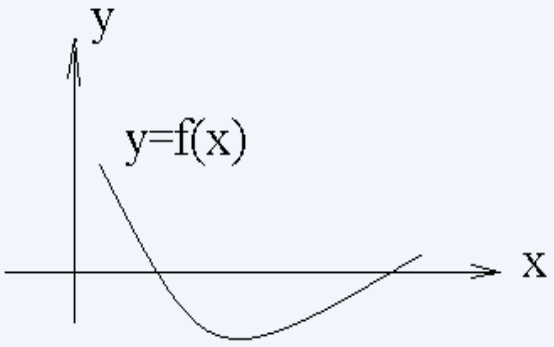
This chapter deals with finding solutions of algebraic and transcendental equations of either of the forms

$$f(x) = 0 \qquad \text{or} \qquad f(x) = g(x) \tag{2.1}$$

where we want to solve for the unknown x. An algebraic equation is an equation constructed using the operations of $+, -, \times, \div$, and possibly root taking (radicals). Rational functions and polynomials are examples of algebraic functions. Transcendental equations in comparison are not algebraic. That is, they contain non-algebraic functions and possibly their inverses functions. Equations which contain either trigonometric functions, inverse trigonometric functions, exponential functions, and logarithmic functions are examples of non-algebraic functions which are called transcendental functions. Transcendental functions also include many functions defined by the use of infinite series or integrals.

Graphical Methods

Confronted with equations having one of the above forms and assuming one has access to a graphical calculator or computer that can perform graphics, then one should begin by plotting graphs of the given functions. If the equation to be solved is of the form $f(x) = 0$, then we plot a graph of $y = f(x)$ over some range of x until we find where the curve crosses the $x-$axis. Points where $y = 0$ or $f(x) = 0$ are called the roots of $f(x)$ or the zeros of $f(x)$.



The point (or points) where the given curve $y = f(x)$ crosses the $x-$axis is where $y = 0$. All such points of intersection then represent solutions to the equation $f(x) = 0$.

Example 2-1. (Root of algebraic equation.)
Estimate the solutions of the algebraic equation

$$f(x) = x^3 - \frac{132}{32}x^2 + \frac{28}{32}x + \frac{147}{32} = 0.$$

Solution: We use a computer or calculator and plot a graph of the function $y = f(x)$ and obtain the figure 2-1.

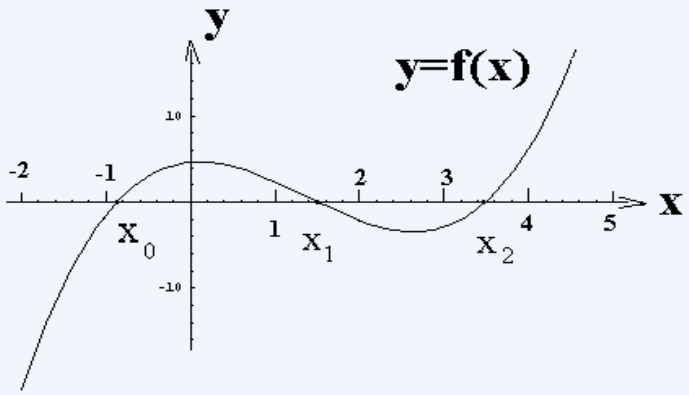


Figure 2-1. Graph of $y = f(x) = x^3 - \frac{132}{32}x^2 + \frac{28}{32}x + \frac{147}{32}$

One can now estimate the solutions of the given equation by determining where the curve crosses the $x-$axis because these are the points where $y = 0$. Examining the graph in figure 2-1 we can place bounds on our estimates x_0, x_1, x_2 of the solutions. One such estimate is given by

$$-1.0 < x_0 < -0.8$$
$$1.4 < x_1 < 1.6$$
$$3.4 < x_2 < 3.6$$

To achieve a better estimate for the roots one can plot three versions of the above graph which have some appropriate scaling in the neighborhood of the roots.

■

Finding values for x where $f(x) = g(x)$ can also be approached using graphics. One can plot graphs of the curves $y = f(x)$ and $y = g(x)$ on the same set of axes and then try to estimate where these curves intersect.

Example 2-2. (Root of transcendental equation.)
Estimate the solutions of the transcendental equation

$$x^3 - \frac{132}{32}x^2 + \frac{28}{32}x + \frac{147}{32} = 5\sin x$$

Solution: We again employ a computer or calculator and plot graphs of the functions

$$y = f(x) = x^3 - \frac{132}{32}x^2 + \frac{28}{32}x + \frac{147}{32} \quad \text{and} \qquad y = g(x) = 5\sin x$$

to obtain the figure 2-2.

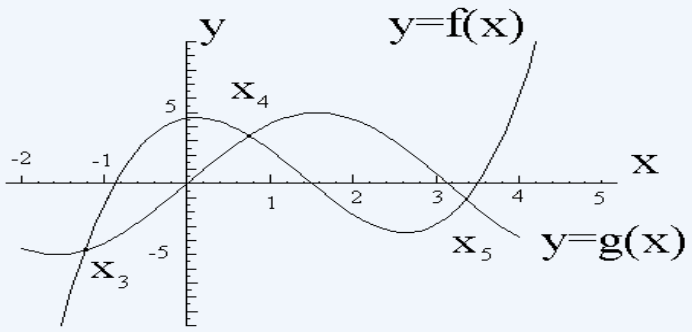


Figure 2-2. Graph of $y = f(x)$ and $y = g(x)$

One can estimate the points where the curve $y = f(x)$ intersects the curve $y = g(x)$. If the curves are plotted to scale on the same set of axes, then one can place bounds on the estimates of the solution. One such set of bounds is given by

$$\begin{aligned} -1.5 < x_3 &< -1.0 \\ 0.5 < x_4 &< 1.0 \\ 3.0 < x_5 &< 3.5 \end{aligned}$$

By plotting these graphs over a finer scale one can obtain better estimates for the solutions.

■

Bisection Method

The bisection method is also known as the method of interval halving. The method assumes that you begin with a continuous function $y = f(x)$ and that you desire to find a root r such that $f(r) = 0$. The method assumes that if you plot a graph of $y = f(x)$, then it is possible to select an interval (a, b) such that at the end points of the interval the values $f(a)$ and $f(b)$ are of opposite sign in which case $f(a)f(b) < 0$. Starting with the above assumptions the intermediate value theorem guarantees that there exists at least one root of the given equation in the interval (a, b). The bisection method is a way of determining the root r to some desired degree of accuracy. The assumed starting situation is illustrated graphically in the figure 2-3.

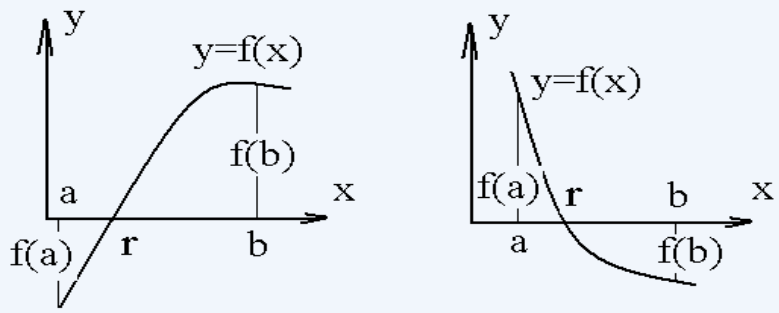


Figure 2-3. Possible starting conditions for the bisection method.

The bisection method generates a sequence of intervals $(a_1, b_1), (a_2, b_2), \ldots, (a_n, b_n)$ which get halved each time. Each interval (a_n, b_n) is determined such that the root r satisfies $a_n < r < b_n$. The bisection method begins by selecting $a_1 = a$ and $b_1 = b$ with $f(a)f(b) < 0$. The midpoint m_1 of the first interval (a_1, b_1) is calculated

$$m_1 = \frac{1}{2}(a_1 + b_1) \tag{2.2}$$

and the curve height $f(m_1)$ is calculated. If $f(m_1) = 0$, then $r = m_1$ is the desired root. If $f(m_1) \neq 0$ then one of the following cases will exist.

Either $f(m_1)f(b_1) < 0$ in which case there is a sign change in the interval (m_1, b_1) or $f(m_1)f(a_1) < 0$ in which case there is a sign change in the interval (a_1, m_1). The new interval (a_2, b_2) is determined from one of these conditions.

(i) If $f(m_1)f(b_1) < 0$, then we select $a_2 = m_1$ and $b_2 = b_1$.

(ii) If $f(m_1)f(a_1) < 0$, then we select $a_2 = a_1$ and $b_2 = m_1$.

Whichever case holds, the root r will lie within the new interval (a_2, b_2) which is one-half the size of the previous interval. The figure 2-4 illustrates some possible scenarios that could result in applying the bisection method to find a root of an equation.

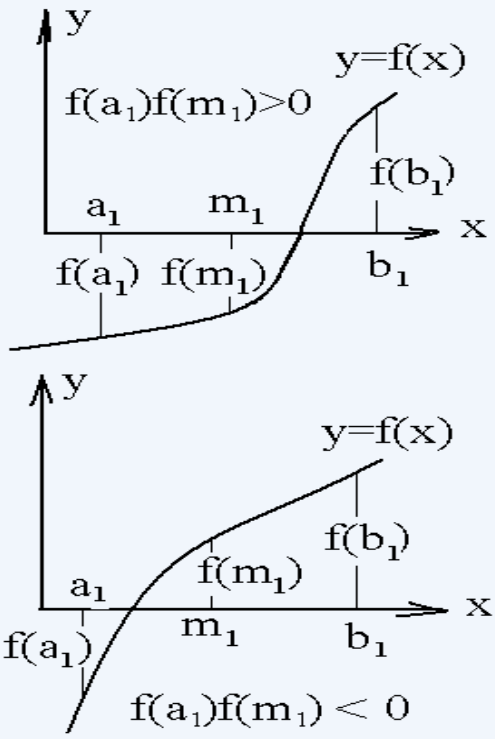


Figure 2-4. Analysis of bisection.

The above process is then repeated as many times as desired to generate new intervals (a_n, b_n) for $n = 3, 4, 5, \ldots$. The bisection method places bounds upon the distance between the n*th* midpoint m_n and the desired root r. One can define the error of approximation after the n*th* bisection as

$$\text{Error} = |r - m_n|. \tag{2.3}$$

This produces the following bounds.

$$\begin{aligned}
\text{After first bisection} \qquad & |r - m_1| < \frac{b-a}{2}, \quad \text{since } r \in (a_1, b_1) \\
\text{After second bisection} \qquad & |r - m_2| < \frac{b-a}{2^2}, \quad \text{since } r \in (a_2, b_2) \\
& \vdots \\
\text{After n}th\text{ bisection} \qquad & |r - m_n| < \frac{b-a}{2^n}, \quad n \geq 1, \qquad \text{since } r \in (a_n, b_n).
\end{aligned}$$

These errors are obtained from the bounds upon the distance between the midpoint m_n and the desired root r. We find the error term associated with the n*th*

step of the iteration procedure for the bisection method will always be less than the initial interval $b-a$ divided by 2^n. If we want the error to be less than some small amount ϵ, then we can require that n be selected such that

$$\text{Error} = |r - m_n| < \frac{b-a}{2^n} < \epsilon. \tag{2.4}$$

The bisection method generates a sequence of midpoint values $\{m_1, m_2, \ldots, m_n, \ldots\}$ used to approximate the true root r. The number of interval halving operations to be performed for a given function $f(x)$ depends upon how accurate you want your solution. The following is a list of some stopping conditions associated with the bisection method.

Stopping Conditions for Bisection Method

(i) If one requires that equation (2.4) be satisfied, then n can be selected as the least integer which satisfies

$$n > \frac{\ln|b-a| - \ln\epsilon}{\ln 2} \tag{2.5}$$

(ii) Given a error ϵ one could continue until $|m_n - m_{n-1}| < \epsilon$. This requires that the two consecutive midpoints be within ϵ of one another.

(iii) One can require that the relative error or percentage error be less than some small amount ϵ. This requires that

$$\frac{|m_n - m_{n-1}|}{|m_n|} < \epsilon \quad \text{or} \quad \frac{|m_n - m_{n-1}|}{|m_n|} \times 100 < \epsilon$$

(iv) The height of the curve $y = f(x)$ is near zero. This requires $|f(m_n)| < \epsilon$ where ϵ is some stopping criteria.

(v) One can arbitrarily select a maximum number of iterations N_{max} and stop the interval halving whenever $n > N_{max}$. One usually selects N_{max} based upon an analysis of equation (2.4).

(vi) The inequality $|r - m_n| \leq |b_n - a_n| < \epsilon$ can be used to define a stopping condition for the error.

Example 2-3. (Bisection method.)

Find the value of x which satisfies $f(x) = xe^x - 2 = 0$.

Solution: We sketch the given function and select $a_1 = 0$ and $b_1 = 1$ with $f(a_1) = -2$ and $f(b_1) = 0.718$. This type of a problem can be easily entered into a spread sheet program which can do the repetitive calculations quickly. Many free spread sheet

programs are available from the internet for those interested. The bisection method produces the following table of values where the error after the nth bisection is less than $E = \frac{b-a}{2^n}$.

Bisection method to solve $f(x) = xe^x - 2 = 0$ **with** $\mid r - m_n \mid < E = \frac{b-a}{2^n}$						
n	a_n	$f(a_n)$	b_n	$m_n = \frac{1}{2}(a_n + b_n)$	$f(m_n)$	E
1	0	-2	1	0.5	-1.1756	0.5
2	0.5	-1.17564	1	0.75	-0.4122	0.25
3	0.75	-0.41225	1	0.875	0.0990	0.125
4	0.75	-0.41225	0.875	0.8125	-0.1690	0.0625
5	0.8125	-0.16900	0.875	0.84375	-0.0382	0.03125
6	0.84375	-0.03822	0.875	0.859375	0.0296	0.015625
7	0.84375	-0.03822	0.859375	0.8515625	-0.0045	0.0078125
8	0.8515625	-0.00453	0.859375	0.85546875	0.0125	0.00390625
9	0.8515625	-0.00453	0.85546875	0.853515625	0.0040	0.001953125
10	0.8515625	-0.00453	0.853515625	0.852539063	-0.0003	0.000976563
11	0.852539063	-0.00029	0.853515625	0.853027344	0.0018	0.000488281
12	0.852539063	-0.00029	0.853027344	0.852783203	0.0008	0.000244141
13	0.852539063	-0.00029	0.852783203	0.852661133	0.0002	0.000122070

Continuing one can achieve the more accurate approximation $r = 0.852605502$.

■

There can be problems in using the bisection method. In addition to the bisection method being slow, there can be the problem that the initial interval (a, b) is selected too large. If this condition occurs, then there exists the possibility that more than one root exists within the initial interval. Observe that if the starting interval contains more than one root, then the bisection method will find only one of the roots. The good thing about the bisection method is that it always works when the setup conditions are satisfied.

Linear Interpolation

The method of linear interpolation is often referred to as the method of false position or the Latin equivalent "regula falsi". It is a method that is sometimes used in the attempt to speed up the bisection method. The method of linear interpolation is illustrated in the figure 2-5.

Given two points $(a_n, f(a_n))$ and $(b_n, f(b_n))$, where $f(a_n)f(b_n) < 0$, then one can construct a straight line through these points. The point-slope formula can be used to find the equation of the line in figure 2-5. One obtains the equation

$$y - f(b_n) = \left(\frac{f(b_n) - f(a_n)}{b_n - a_n}\right)(x - b_n) \qquad (2.6)$$

This line crosses the $x-$axis at the point $(x_n, 0)$ where

$$x_n = b_n - \left(\frac{f(b_n)}{f(b_n) - f(a_n)}\right)(b_n - a_n).$$

We use the point x_n in place of the midpoint m_n of the bisection method. That is, each iteration begins with end points $(a_n, f(a_n))$ and $(b_n, f(b_n))$ where $f(a_n)$ and $f(b_n)$ are of opposite sign. These points produce a straight line which determines a point x_n by linear interpolation.

A new interval (a_{n+1}, b_{n+1}) is determined by the same procedure used in the bisection method. We calculate $f(x_n)$ and test the sign of $f(x_n)f(a_n)$ in order to determine the new interval (a_{n+1}, b_{n+1}) which contains the desired root r. The method of linear interpolation sometimes has the problem of a slow one-sided approach to the root as illustrated in the figure 2-6.

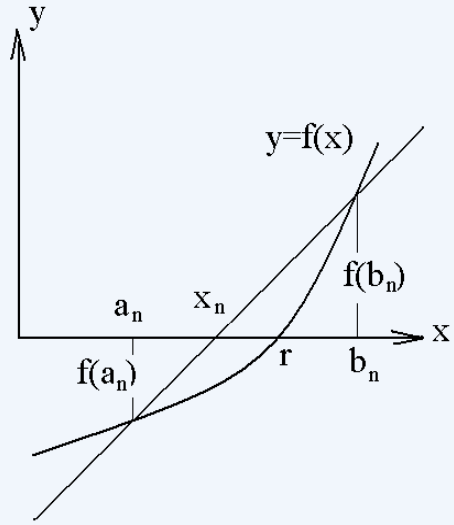


Figure 2-5.
Linear interpolation.

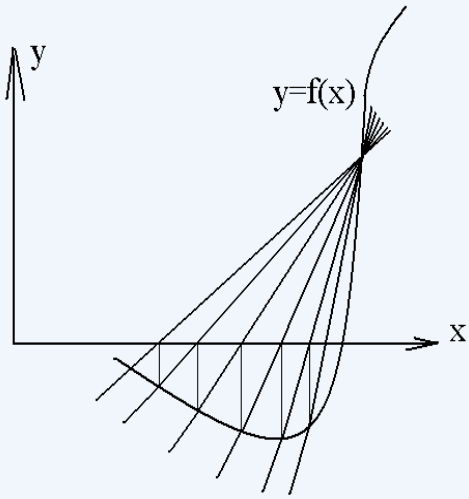


Figure 2-6.
Slow one-sided approach.

Iterative Methods

One can write equations of the form $f(x) = 0$ in the alternative form $x = g(x)$ and then one can define an iterative sequence

$$x_{n+1} = g(x_n) \quad \text{for} \quad n = 0, 1, 2, 3, \ldots \tag{2.7}$$

which can be interpreted as mapping a point x_n to a new point x_{n+1}. One starts with an initial guess x_0 to the solution of $x = g(x)$ and calculates $x_1 = g(x_0)$. The iterative method continues with repeated substitutions into the $g(x)$ function to obtain the values

$$\begin{aligned} x_2 &= g(x_1) \\ x_3 &= g(x_2) \\ &\vdots \\ x_n &= g(x_{n-1}) \\ x_{n+1} &= g(x_n) \end{aligned}$$

If the sequence of values $\{x_n\}_{n=1}^{\infty}$ converges to r, then

$$\lim_{n\to\infty} x_{n+1} = r = \lim_{n\to\infty} g(x_n) = g(r)$$

and r is called a fixed point of the mapping. Convergence of the iterative processes is based upon the concept of a contraction mapping. In general, a mapping $x_{n+1} = g(x_n)$ is called a contraction mapping if the following conditions are satisfied.

1. The function $g(x)$ maps all point in a set S_n into a subset S_{n+1} of S_n so that one can write $S_{n+1} \subset S_n$.
2. For $x_n, y_n \in S_n$, with $x_{n+1} = g(x_n)$ and $y_{n+1} = g(y_n)$ both members of the set S_{n+1}, the distance between y_{n+1} and x_{n+1} must be less than the distance between y_n and x_n. This can be expressed

$$| y_{n+1} - x_{n+1} | \leq K | y_n - x_n |$$

where K is some constant satisfying $0 \leq K < 1$.

That is, the distance between any two points x_n and y_n belonging to a set S_n is always greater than the distance between the image points x_{n+1} and y_{n+1} belong to the image subset S_{n+1}. The representation of a contraction mapping

is illustrated in the figure 2-7 which gives an image showing points from one set being mapped to a smaller set. This is the idea behind a contraction mapping. Each mapping gives a smaller and smaller image set which eventually contracts to a limit point r where $r = g(r)$. This idea can be applied to more general types of mappings.

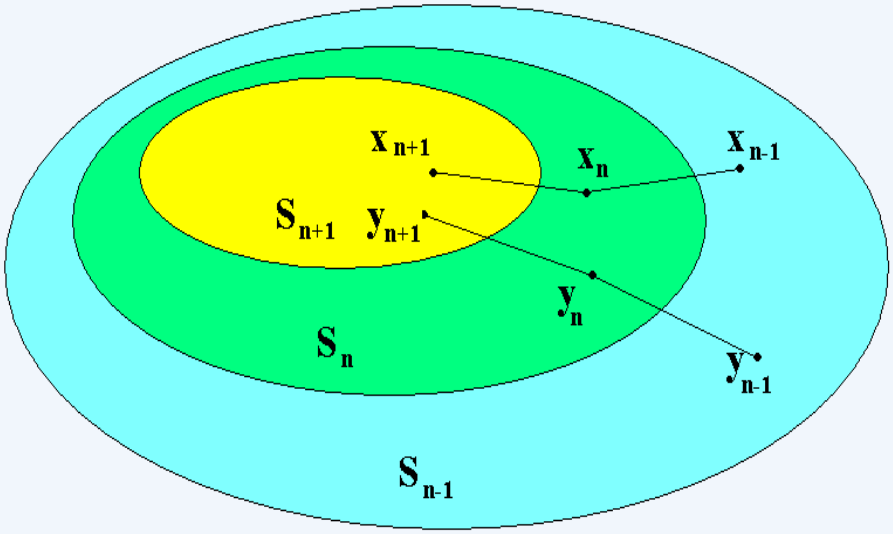


Figure 2-7. Contraction mapping.

In one-dimension, assume that the iterative sequence

$$x_{n+1} = g(x_n) \tag{2.8}$$

converges to a limit r such that

$$r = g(r). \tag{2.9}$$

Subtract the equation (2.9) from the equation (2.8) and write

$$x_{n+1} - r = g(x_n) - g(r) = \left[\frac{g(x_n) - g(r)}{x_n - r}\right](x_n - r). \tag{2.10}$$

The mean value theorem can now be employed to express the bracketed term in equation (2.10) in terms of a derivative so that

$$\left[\frac{g(x_n) - g(r)}{x_n - r}\right] = g'(\xi_n) \qquad \text{for} \qquad x_n < \xi_n < r. \tag{2.11}$$