$ax^2+bx+c=0$ $\qquad$ $x=\frac{-b\pm\sqrt{b^2-4ac}}{2a}$ $\quad (a\neq 0)$

$\sin x - \frac{1}{2}x^3+3x=0$

$ax+b=0 \longmapsto x=-b/a \qquad (a\neq 0)$

Module 2 – Root finding methods

Lesson goals

1. Understanding what roots problems are and where they occur in engineering and science.
2. Knowing how to solve a roots problem with the bisection method.
3. Knowing how to estimate the error of bisection and why it differs from error estimates for other types of root-location algorithms.
4. Understanding false position and how it differs from bisection.
5. Understanding the distinction between accuracy and precision.

Introduction

Assume that $f(x)$ is a continuous function in its domain. The aim is to find value(s) for x so that

$$f(x)=0$$

In general, there is no closed form formula for the roots (zeros) of the function f. In specific cases, we know how to find the root(s) of a function, such as linear and quadratic functions. To recall, if the function $f(x)=ax^2+bx+c$ has nonnegative discriminant b^2-4ac, then it has real roots:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

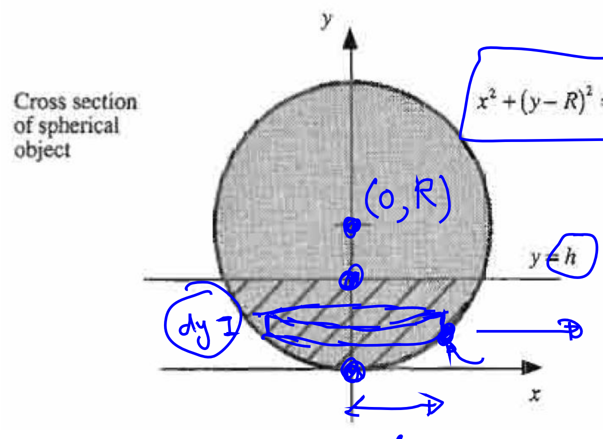
As there are no closed form formula for the given function in general, we need to develop algorithms so that we approximate the roots. The root finding techniques are designed to handle this matter.

Why do we need to learn root finding techniques in engineering and sciences? To answer this question, here is an example.

Example. Suppose we want to determine how far a spherical object of radius R will sink into a fluid such as water or oil. According to Archimedes' principal, the object will sink to the depth at which the weight of water displaced by the object equals the weight of the object. Now, the weight of the object is the product of its mass, m, and the acceleration due to gravity, g. If we assume that the object has a constant mass density, ρ_0, then $m = \frac{4}{3}\pi R^3 \rho_0$ and

$$\text{Weight of object} = \frac{4}{3}\pi R^3 \rho_0 g \quad = \text{Weight of displaced fluid.} \qquad (1)$$

$$\text{Weight} = \underbrace{\text{mass}}\times g$$
$$= \rho_0 \frac{4}{3}\pi R^3 g$$



$$= \rho_f V_d g$$

$$dV = \pi x^2 dy$$
$$= \pi(R^2 - (y-R)^2)dy$$
$$= \pi(2Ry - y^2)dy$$

Assuming the fluid has density ρ_f and V_d is the volume of fluid displaced by the object, then

$$\text{Weight of displaced fluid} = \rho_f V_d g \qquad (2)$$

To complete the specification of the problem, we need to determine V_d. Suppose the object sinks to the depth h. Considering the geometry shown in the diagram above and applying some basic calculus (volume by slicing to be exact), we find that

$$V_d = \int_0^h dV = \pi \int_0^h (2Ry - y^2)dy = \pi\left(y^2R - \frac{1}{3}y^3\right)\Big]_0^h = \pi\left(h^2R - \frac{1}{3}h^3\right)$$

$$V_d = \pi \int_0^h x^2 dy = \pi \int_0^h (2Ry - y^2) dy = \pi h^2 \left(R - \frac{h}{3}\right) \qquad (3)$$

Substituting (3) into (2) and equating the resulting expression with (1) yields

$$\frac{4}{3}\pi R^3 \rho_0 g = \pi h^2 \left(R - \frac{h}{3}\right) \rho_f g$$

After some algebraic simplification, it becomes

Solve for h

$$\boxed{\frac{\rho_f}{3} h^3 - R\rho_f h^2 + \frac{4}{3} R^3 \rho_0 = 0.}$$

Therefore, given R, ρ_0, and ρ_f, the depth to which the object sinks, is determined by solving this equation for h.

$f(x) = 0$

$f(p) = 0$

Multiplicity of a root: A root p of the equation $f(x) = 0$ is said to be a root of multiplicity m if f can be written in the form $f(x) = (x - p)^m q(x)$, where $q(x) \neq 0$. A root with multiplicity one is called simple root.

$\longmapsto q(p) \neq 0$

Example. For the equation

$f(x)$ is divisible by $(x-p)$

$$\boxed{x^6 + x^5 - 12x^4 + 2x^3 + 41x^2 - 51x + 18 = 0}$$

Can be written as

$x^6+x^5-12x^4+2x^3+41x^2-51x+18 = (x-1)\,g(x)$

$f(x) = (x-p)\,g(x)$

$$(x - 1)^3 (x + 3)^2 (x - 2) = 0, \quad x = 1 \text{ (mult. 3)}, x = -3 \text{(mult. 2)}, x = 2 \text{ (simple)}$$

if $g(p) \neq 0 \longmapsto p$ is called a simple root of $f(x)$.

if $g(p) = 0 \longmapsto p$ is called a repeated root of $f(x)$.

Simple criteria: For the continuous and differentiable function f, the equation $f(x) = 0$ has a root of multiplicity m at $x = p$ if and only if

$\longmapsto f(x) = (x-p)(x-p)\,h(x)$

$= (x-p)^2\,h(x)$

$$f(p) = f'(p) = \cdots = f^{(m-1)}(p) = 0, \text{and } f^{(m)}(p) \neq 0$$

Example. Show that the function $f(x) = 2x + \ln\left(\frac{1-x}{1+x}\right)$ has a root of multiplicity 3 at $x = 0$.

$f(0) = 0 + \ln(1) = 0$

$$f'(x) = 2 + \frac{\frac{-2}{(1+x)^2}}{\frac{1-x}{1+x}} = 2 - \frac{2}{(1-x)(1+x)} = 2 - \frac{2}{1-x^2} \longmapsto f'(0) = 0$$

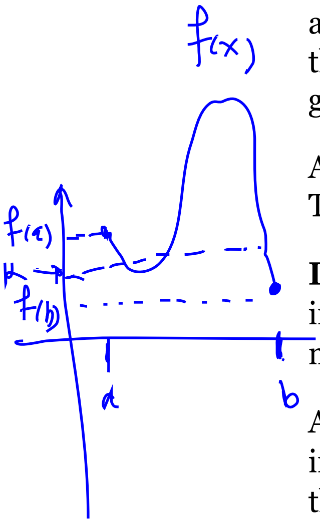
$$f''(x) = \frac{-4x}{(1-x^2)^2} \longmapsto f''(0) = 0$$

$$f'''(x) = \frac{-4(1-x^2) - 16x^2}{(1-x^2)^3} \longmapsto f'''(0) = \frac{-4}{1} = -4 \neq 0$$

multiplicity order of $x=0$ is $\boxed{3}$

Numerical root finding methods

These techniques are generally divided into two categories: Simple bracketing methods and iterative schemes. The bracketing methods determine small intervals that contain the root. In the iterative methods a sequence of approximate values for the root is generated that converges to the root.

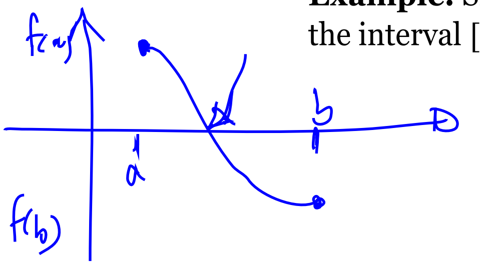


All simple bracketing (enclosure) methods are based on the Intermediate Value Theorem.

Intermediate Value Theorem: Let $f(x)$ be a real and continuous function in the interval $[a, b]$ and k be any real value between $f(a)$ and $f(b)$. Then, there exists a real number $c \in (a, b)$ such that $f(c) = k$.

As a consequence of Intermediate Value Theorem, if $f(x)$ is real and continuous in the interval from x_l to x_u and $f(x_l)$ and $f(x_u)$ have opposite signs, that is, $f(x_l)f(x_u) < 0$, then there is at least one real root between x_l and x_u.

Example. Show that the function $f(x) = \sin(10x) + \cos(3x)$ has at least one root in the interval $[4.5\text{rad}, 5\text{rad}]$. Graph this function in Matlab to see this fact!

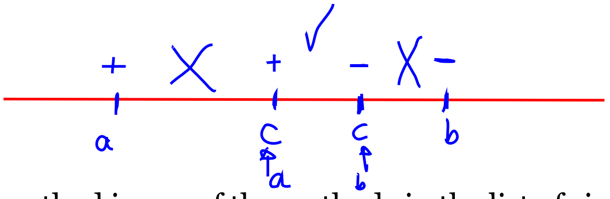


$$\left\{ \begin{array}{l} f \text{ is continuous} \\ f(a)\,f(b) < 0 \end{array} \right. \longmapsto$$ by IVT, there is at least one root in the interval (a,b)

$[a,b] = [4.5, 5]$

$f(x) = \sin(10x) + \cos(3x)$

$$\longmapsto \left\{ \begin{array}{l} f \text{ is a cont. function} \\ f(4.5) = 1.446 \\ f(5) = -1.022 \end{array} \right. \longmapsto$$ by IVT, there is at least one root in this interval.



The bisection method is one of the methods in the list of simple enclosure methods.

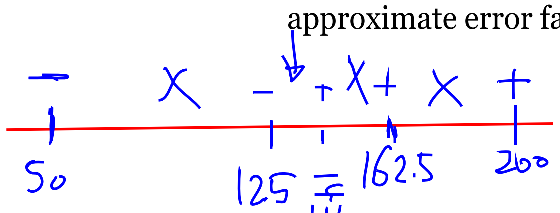
The bisection method

Assume that the Intermediate Value Theorem has been applied and an interval containing a zero x^* of the given continuous function is located.

In the bisection method the interval is always divided in half and the function value at the midpoint is evaluated. The location of the root is then determined as lying within the subinterval where the sign change occurs.

For notational convenience, assume that (a_r, b_r) is the enclosing interval during the r −th interaction, and $x_r = \frac{a_r+b_r}{2}$ is the midpoint of this interval. If x_r is an accurate enough approximation, let say $\left|\frac{x_r-x^*}{x^*}\right| < \varepsilon_s$, or $\left|\frac{x_r-x_{r-1}}{x_{r-1}}\right| < \varepsilon_s$, for the prespecified accuracy ε_s, then the process is terminated. Otherwise, the Intermediate Value Theorem is invoked to determine which of two subintervals, (a_r, x_r) or (x_r, b_r), contains the root and becomes (a_{r+1}, b_{r+1}).

Example. Use the bisection method to solve the following equation so that approximate error falls below a stopping criterion of $\varepsilon_s = 0.5\%$.



$$\boxed{f(m) = \sqrt{\frac{gm}{c_d}}\tanh\left(\sqrt{\frac{gc_d}{m}}\,t\right) - v = 0}$$

where $t = 4, g = 9.81, c_d = 0.25$, and $v = 36$.

Solution. We can see that the function changes sign between values of 50 and 200. So,

$$x_1 = \frac{50 + 200}{2} = 125$$

For the second iteration, the root lies between 125 and 200. So, the second estimate is:

$$x_2 = \frac{125 + 200}{2} = 162.5$$

And so on ...

Note that, at each iteration, the error is computed to see if it satisfies the stopping criterion. The results are summarized in the following table.

Iteration	a_r	b_r	x_r	$\lvert\varepsilon_a\rvert$ (%)	$\lvert\varepsilon_t\rvert$ (%)
1	50	200	125		12.43
2	125	200	162.5	23.08	13.85
3	125	162.5	143.75	13.04	0.71
4	125	143.75	134.375	6.98	5.86
5	134.375	143.75	139.0625	3.37	2.58
6	139.0625	143.75	141.4063	1.66	0.93
7	141.4063	143.75	142.5781	0.82	0.11
→ 8	142.5781	143.75	143.1641	0.41	0.30

While (1)

$\varepsilon_s = \frac{1}{2} \times 10^{-3}$ ✓

Error bound: Let $f(x)$ be continuous in the interval $[a, b]$ and suppose that $f(a)f(b) < 0$. The bisection method generates a sequence of approximations $\{x_r\}$ which converges to the root $x^* \in (a, b)$ with the property

$$|x_r - x^*| \le \frac{b-a}{2^{n \to r}} = \text{error upper bound}$$

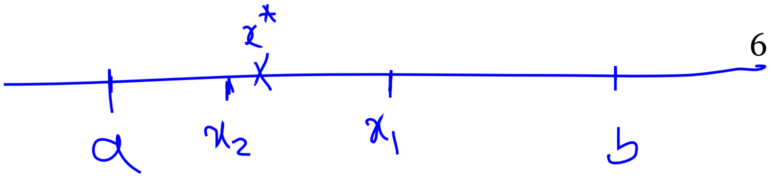
In the bisection method, the number of iterations required to attain an absolute error can be computed *a priori*. If ε_s is the desired error, then the number of iterations can be obtained by solving the following inequality for n:

$$\frac{b-a}{2^n} \le \varepsilon_s.$$

$|x_1 - x^*| \le \frac{1}{2}(b-a)$

$|x_2 - x^*| \le \frac{1}{2^2}(b-a)$

$|x_3 - x^*| \le \frac{1}{2^3}(b-a)$



Example. Find the minimum number of iterations that requires to get an estimate of the root with prespecified error $\varepsilon_s = 0.5\%$.

$a=1, \quad b=4$

$$\mathcal{E}_s = 0.5\% = 0.005$$

We make the upper bound for the error to be less than $\mathcal{E}_s$.

$$\frac{b-a}{2^r} < 0.005 \mapsto \frac{4-1}{2^r} < 0.005 \mapsto 2^r > \frac{3}{0.005}$$

$$\to r > \log_2 3/0.005 = 9.23 \mapsto \underline{\underline{r=10}}$$

Example (try it yourself). Consider the following equation:

$$\tan(\pi x) - x - 6 = 0$$

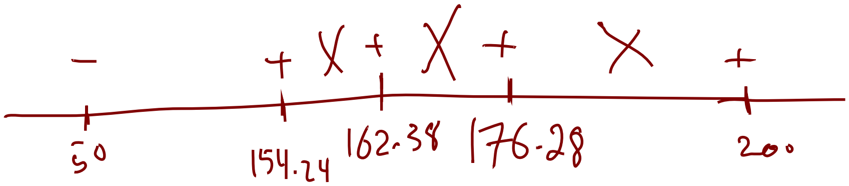
Approximate the smallest positive root that lies on the interval (0.4, 0.48) with $\varepsilon_s = 5 \times 10^{-5}$.

# of iterations needed to get the accuracy according to $\mathcal{E}_s = 5\times10^{-5}$

$a=0.4$
$b=0.48$

$$\longmapsto \frac{b-a}{2^r} < \mathcal{E}_s \mapsto \frac{0.08}{2^r} < 5\times10^{-5} \mapsto 2^r > \frac{0.08}{5}\times10^{5}$$

$$r > \log_2\left(\frac{0.08}{5}\times10^{5}\right) = 10.64 \mapsto r \geq 11$$



False position. It is another bracketing method. It is very similar to bisection method. Rather than bisecting the interval, it locates the root by joining $f(x_l)$ and $f(x_u)$ with a straight line (see the following Figure). The intersection of this line with the x axis represents an improved estimate of the root. The formula is:

$$x_r = x_u - \frac{f(x_u)(x_l - x_u)}{f(x_l) - f(x_u)}$$

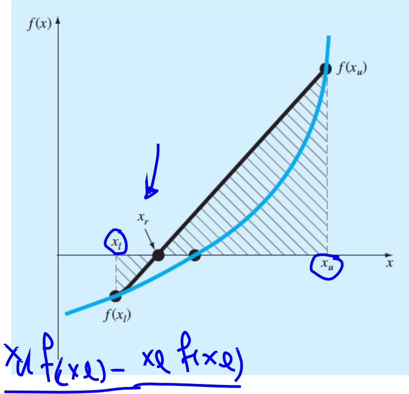
The value of x_r then replaces whichever of the two initial guesses, x_l or x_u, yields a function value with the same sign as $f(x_r)$. The process is repeated until the root is estimated adequately.

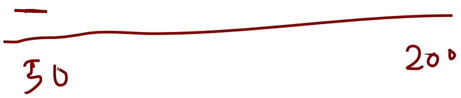
$$\left.\begin{cases}(x_l, f(x_l))\\(x_u, f(x_u))\end{cases}\right\} \rightarrow m = \frac{f(x_u) - f(x_l)}{x_u - x_l}$$

$$y - f(x_l) = \frac{f(x_u) - f(x_l)}{x_u - x_l}(x - x_l)$$

$$0 - f(x_l) = \frac{f(x_u) - f(x_u)}{x_u - x_l}(x - x_l)$$

$$\rightarrow x - x_l = -\frac{x_u - x_l}{f(x_u) - f(x_l)} f(x_l) = x_l - \frac{x_u f(x_l) - x_l f(x_l)}{f(x_u) - f(x_l)}$$





Example. Consider the equation:

$$f(m) = \sqrt{\frac{gm}{c_d}} \tanh\left(\sqrt{\frac{gc_d}{m}}\, t\right) - v = 0$$

where $t = 4, g = 9.81, c_d = 0.25$, and $v = 36$. Perform *two iterations* of the false-position method to approximate error the root.

Solution. 1st iteration:

$$x_l = 50, \qquad f(x_l) = -4.579387$$

$$x_u = 200, \qquad f(x_u) = 0.860291$$

$$x_1 = 200 - \frac{0.860291(50 - 200)}{-4.579387 - 0.860291} = 176.2773$$

2nd iteration:

$$x_l = 50, \qquad f(x_l) = -4.579387$$

$$x_u = 176.2773, \qquad f(x_u) = 0.566174$$

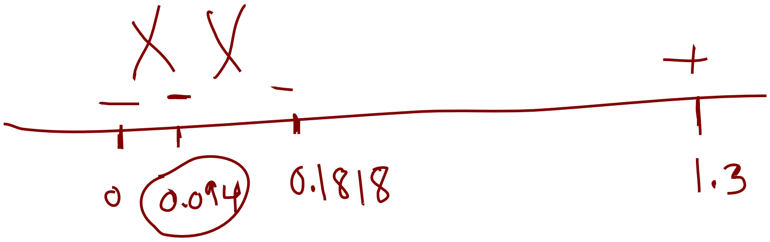
$$x_2 = 176.2773 - \frac{0.566174(50 - 176.2773)}{-4.579387 - 0.566174} = \underline{162.3828}$$

Compute the third iteration (practice at home)

Example (try it yourself). Use bisection and false position to locate the root of

$$f(x) = x^{10} - 1$$

between $x = 0$ and 1.3.



$g(x)$

$g(p)=p$

$f(p)=0$ $\quad f(x) = 0 \longmapsto \underbrace{x=g(x)}$

$p=g(p)$

Simple fixed-point iteration

The fixed-point (successive substitution) method is one of the open methods that employs a formula to predict the root. Such a formula can be developed by rearranging the function $f(x) = 0$ so that x is on the left-hand side of the equation:

$$x = g(x)$$

This transformation can be done either by algebraic manipulation or by simply adding x to both sides of the original equation.

Given an initial guess at the root x_i, this equation can be used to compute a new estimate x_{i+1} as expressed by the iterative formula

x_0

$$✓ \boxed{x_{i+1} = g(x_i)}$$

As with many other iterative formulas, the approximate error for this equation can be determined using the error estimator:

$$\varepsilon_a = \left|\frac{x_{i+1} - x_i}{x_{i+1}}\right| \times 100\%. < \varepsilon_s$$

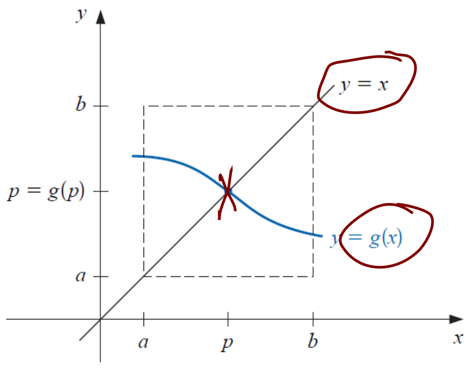
What is fixed point of a function? A fixed point of a function g is any real number p for which $g(p) = p$; that is, whose location is fixed by g.

Which function has a fixed point? Let g be a continuous function on the interval $[a, b]$ and $g: [a, b] \to [a, b]$. Then, g has a fixed point. Furthermore, if g is differentiable on (a, b), and there is a positive constant $k < 1$ so that $|g'(x)| \le k < 1$, for all $x \in (a, b)$, then the fixed point is unique.

$$\boxed{g: \widetilde{[a,b]} \longrightarrow \widetilde{[a,b]}} ✓$$

$|g'(x)| \le K < 1$

on $[a,b]$

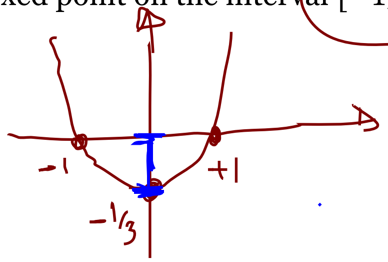


$g([a,b]) \subseteq [a,b]$ ✓

$|g'(x)| \le K < 1$

Example. Show that $g(x) = \frac{x^2-1}{3}$ has a unique fixed point on the interval $[-1, 1]$.

$$g(x) = \frac{x^2-1}{3} = \frac{1}{3}x^2 - \frac{1}{3}$$



① What is the range of $g(x)$ on $[-1,1]$?

$$R_g = [-\tfrac{1}{3}, 0] \subseteq [-1,1]$$

So, $g(x)$ has at least one fixed-point

② $g'(x) = \frac{2}{3}x \mapsto |g'(x)| \leq \frac{2}{3} = K < 1 \longrightarrow$ The fixed point is unique

on $[-1,1]$

Example (try it yourself!) Show that the function $g(x) = 3^{-x}$ has a unique fixed point on the interval $[0, 1]$.

$g(x)$

$h(x)$

$F(x) = g(x) - h(x) = 0 \longrightarrow g(x) = h(x)$

$$\varepsilon_t = \frac{|x_n - x^*|}{|x^*|} \quad - \qquad \varepsilon_a = \frac{|x_n - x_{n-1}|}{|x_{n-1}|} = \frac{|x_n - x_{n-1}|}{|x_n|}$$

Procedure of the fixed-point method: Assume that $f(x) = 0$ has a simple root x^* on $[a, b]$.

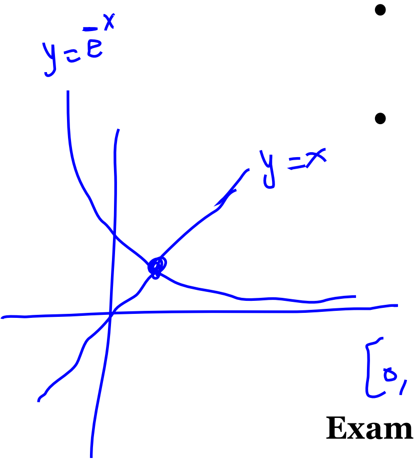
$x = g(x)$

- Rewrite $f(x) = 0$ in the form $x = g(x)$; choose a proper $g(x)$ so that x^* is its unique fixed point
- For the initial point $x_0 \in (a, b)$, generate the sequence of points $\{x_n\}$ using the following iterative scheme:

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

$$\frac{|x_{n+1} - x_n|}{|x_n|} < \varepsilon_s$$

Under certain conditions, this sequence converges to x^*.



$[0, 1]$

Example. Use simple fixed-point iteration to locate the root of $f(x) = e^{-x} - x$.

Consider $[0,1]$. The functi $f(x)$ has a root on $[0,1]$.

Solution. The iterative scheme can be written as:

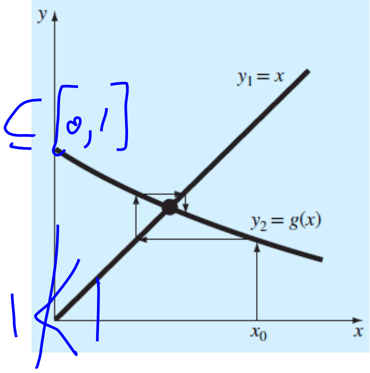
$$x_{i+1} = e^{-x_i}$$

$f(0) = 1 - 0 = 1 > 0$
$f(1) = e^{-1} - 1 < 0$
$\longrightarrow$ by IVT the funct has at least one root

$f(x) = 0$
$\rightsquigarrow e^{-x} - x = 0$
$\longrightarrow \underbrace{e^{-x}}_{g(x)} = \underline{x}$

Starting from $x_0 = 0$, we have the following table of points.

$$\begin{cases} g([0,1]) = [\frac{1}{e}, 1] \subseteq [0,1] \\ g'(x) = -e^{-x} \\ |g'(x)| = |-e^{-x}| = 1 \end{cases}$$



i	x_i	$\lvert\varepsilon_a\rvert$, %	$\lvert\varepsilon_t\rvert$, %	$\lvert\varepsilon_t\rvert_i / \lvert\varepsilon_t\rvert_{i-1}$
0	0.0000		100.000	
1	1.0000	100.000	76.322	0.763
2	0.3679	171.828	35.135	0.460
3	0.6922	46.854	22.050	0.628
4	0.5005	38.309	11.755	0.533
5	0.6062	17.447	6.894	0.586
6	0.5454	11.157	3.835	0.556
7	0.5796	5.903	2.199	0.573
8	0.5601	3.481	1.239	0.564
9	0.5711	1.931	0.705	0.569
10	0.5649	1.109	0.399	0.566

In order to have both conditions for unique fixed-point satisfied, we change the interval $[0,1]$ to a smaller one.

$$\boxed{I_1 = [0.1, 1]} \longrightarrow \begin{cases} g([0.1, 1]) = [\frac{1}{e}, e^{-0.1}] = [\frac{1}{e}, 0.9] \subseteq [0.1, 1] \\ |g'(x)| = |-e^{-x}| = e^{-x} \leq 0.9 = k < 1 \end{cases}$$

$$f(x) = e^{-x} - x \longmapsto x = e^{-x} = g(x)$$

$$x_0 = 0, \quad x_{n+1} = g(x_n) \longmapsto x_{n+1} = e^{-x_n}$$

x	error (abs)
$x_0 = 0$	
$x_1 = 1$	1
$x_2 = 0.3679$	0.6321
0.6922	0.3243
0.5005	0.1917

Example (try it yourself). The equation $x^3 + 4x^2 - 10 = 0$ has a root in the interval [1,2]. Use fixed point iterative method with $x_0 = 1.5$ and $g(x) = x - \frac{x^3+4x^2-10}{3x^2+8x}$ to approximate the root that is exact up to 7 decimal places, i.e. $\varepsilon_s = 0.5 \times 10^{-7}$.

$$f(x) = x^3 + 4x^2 - 10 = 0 \longmapsto x = g(x)$$

$+x$ $+x$

$$\underbrace{x^3 + 4x^2 + x - 10}_{g_1(x)} = x$$

$$g_1([1,2]) \not\subseteq [1,2]$$

$$\underbrace{x^3 + 4x^2} - 10 = 0$$

$$x^2(x+4) - 10 = 0$$

$$x^2 = \frac{10}{x+4} \longmapsto x = \sqrt{\frac{10}{x+4}} \quad , g_2$$

$$x = g(x) \to \cancel{x} = \cancel{x} - \frac{x^3+4x^2-10}{3x^2+8x}$$

$$\to x^3 - 4x^2 - 10 = 0$$

Convergence Theorem. Let g be a continuous function on the interval $[a, b]$ and $g: [a, b] \to [a, b]$. Furthermore, suppose that g is differentiable on (a, b), and there is a positive constant $k < 1$ so that $|g'(x)| \le k < 1$, for all $x \in (a, b)$. Then the sequence $\{x_n\}$ generated by $x_{n+1} = g(x_n)$ converges to the fixed point x^* for any $x_0 \in (a, b)$, and

$$|x_n - x^*| \le \frac{k^n}{1-k} |x_1 - x_0|$$

$$x_1 = g(x_0)$$

Note 1. As it is seen from this theorem, the speed of convergence depends on the value of k. The smaller value of k results in fast convergence rate.

Note 2. The condition $|g'(x)| \le k < 1$ is a sufficient condition (not necessary) for the convergency. As an example, for the root of $2x^2 - x = 0$ on the interval $[0, 1]$, the iterative scheme $x_{n+1} = g(x_n) = 2x_n^2$ with $x_0 = 0.5$ generates a sequence of points that converges to the root of this equation while $g(x) = 2x^2$ does not satisfy $|g'(x)| \le k < 1$.

$g_1([0.51, 0.7]) \subseteq [0.51, 0.7]$

Example. For the root of $f(x) = x^2 + x - 1$ on the interval $\overbrace{[0.51, 0.7]}$, consider the following three choices for $g(x)$. ✓

1. $g_1(x) = 1 - x^2$
2. $g_2(x) = \sqrt{1-x}$ $\qquad g_2([0.51, 0.7]) \subseteq [0.51, 0.7]$
3. $g_3(x) = \frac{1}{x+1}$ $\qquad g_3([0.51, 0.7]) \subseteq [0.51, 0.7]$

Which of these functions is a proper choice for the fixed-point method? Why?

$$g_1'(x) = -2x \mapsto |g_1'(x)| = 2x \leq 1.4 = k \not< 1$$

$$g_2'(x) = \frac{-1}{2\sqrt{1-x}} \mapsto |g_2'(x)| = \frac{1}{2\sqrt{1-x}} \leq 0.913 = k < 1 \ ✓$$

$$g_3'(x) = \frac{-1}{(x+1)^2} \mapsto |g_3'(x)| = \frac{1}{(x+1)^2} \leq 0.346 = k < 1$$

$g_3(x)$ is a good choice

Upper bounds for the absolute error: Assume that x^* is the root of the function $f(x)$ on $[a, b]$ and $x_0 \in (a, b)$. Let $g(x)$ satisfy the requirements of the convergence theorem for the fixed-point method. Then,

$$\boxed{|x_n - x^*| \leq k^n \max\{x_0 - a, b - x_0\}} \ ✓$$

and

$$|x_n - x^*| \leq \frac{k^n}{1-k}|x_0 - x_1| \ ✓$$

Ex. Assume that we use $g_3(x)$ in the previous example with $x_0 = 0.6$. How many iterations are needed to get an estimate with the ~~accuracy~~ abs. error less than 10^{-5}.

$$k^n \max\{0.6 - 0.51, 0.7 - 0.6\} < 10^{-5} \to (0.346)^n \cdot 0.1 < 10^{-5}$$

$$(0.346)^n < \frac{10^{-5}}{0.1} = 10^{-4} \to n > \log_{0.346} 10^{-4} = \frac{\ln(10^{-4})}{\ln(0.346)} = 8.6 \uparrow \boxed{9}$$

Rate of convergence: The rate of convergence is a measure of how fast the difference between the solution point and its estimates goes to zero. Assume that $\boxed{\{x_n\}}$ is the sequence of points generated by an algorithm that converges to $\underline{\underline{x^*}}$. Then, the rate of convergence is

- Linear if there is a constant $r \in (0,1)$ so that $\frac{\overbrace{|x_{n+1}-x^*|}^{e_{n+1}}}{\underbrace{|x_n-x^*|}_{e_n}} \le \boxed{r}$. For example, $x_n = 1 + \left(\frac{1}{2}\right)^n$ converges linearly to $x^* = 1$.
- Superlinear if $\lim_{n\to\infty} \frac{\overbrace{|x_{n+1}-x^*|}^{e_{n+1}}}{|x_n-x^*|} = 0$. For example, $\boxed{x_n = 1 + \left(\frac{1}{n}\right)^n}$ converges superlinearly to $x^* = \underline{\underline{1}}$. e
- Quadratic if there is a constant $M > 0$ so that $\frac{\overbrace{|x_{n+1}-x^*|}^{e_{n+1}}}{\underbrace{|x_n-x^*|^{\boxed{2}}}_{e_n^2}} \le \boxed{M}$. For example, $x_n = 1 + \left(\frac{1}{n}\right)^{2^n}$ converges quadratically to $x^* = 1$.
- $p > 0$ if there is a constant $M > 0$ so that $\frac{|x_{n+1}-x^*|}{|x_n-x^*|^p} \le M$.

Note. It is shown that the convergence rate of the fixed-point method is *at least linear*.

$$x_n = \left(\frac{8}{9}\right)^n \to 0$$

$$e_{n+1} = \left|\left(\frac{8}{9}\right)^{n+1} - 0\right| = \left(\frac{8}{9}\right)^{n+1}$$

$$e_n = \left|\left(\frac{8}{9}\right)^n - 0\right| = \left(\frac{8}{9}\right)^n$$

$$\frac{e_{n+1}}{e_n} = \frac{8}{9}$$

$$e_{n+1} = |x_{n+1} - x^*| = \left(\frac{1}{n+1}\right)^{n+1}$$

$$e_n = |x_n - x^*| = \left(\frac{1}{n}\right)^n$$

$$\lim_{n\to\infty} \frac{e_{n+1}}{e_n} = \lim_{n\to\infty} \frac{\left(\frac{1}{n+1}\right)^{n+1}}{\left(\frac{1}{n}\right)^n} = \lim_{n\to\infty} \frac{n^n}{(n+1)^{n+1}} = \lim_{n\to\infty} \boxed{\frac{n^n}{(n+1)^n}} \times \frac{1}{n+1} = \frac{1}{e} \times 0 = 0$$

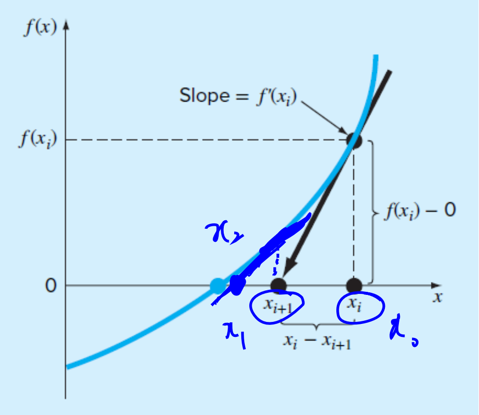
$$e = \lim_{n\to\infty}\left(1 + \frac{1}{n}\right)^n$$

Newton-Raphson method

Newton's (or the Newton-Raphson) method is one of the most powerful and well-known numerical methods for solving a root-finding problem.

If the initial guess for the root is x_i, then the point where the tangent line at the point $(x_i, f(x_i))$ crosses the x axis usually represents an improved estimate of the root. Therefore, the iterative scheme of the Newton's method is:

$$\boxed{x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}}$$



Example. Use the Newton-Raphson method to estimate the root of $f(x) = e^{-x} - x$ employing an initial guess of $x_0 = 0$.

Solution. The Newton's iterative scheme is:

$$f(x) = e^{-x} - x$$

$$f'(x) = -e^{-x} - 1$$

$$x - \frac{f(x)}{f'(x)} = x - \frac{e^{-x} - x}{-e^{-x} - 1}$$

$$\boxed{x_{i+1} = x_i - \frac{e^{-x_i} - x_i}{-e^{-x_i} - 1}}$$

Starting with $x_0 = 0$, we have the following table.

$$x_0 = 0$$

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = 0.5$$

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = 0.56631$$

i	x_i	$\lvert\varepsilon_t\rvert$, %
0	0	100
1	0.500000000	11.8
2	0.566311003	0.147
3	0.567143165	0.0000220
4	0.567143290	$<10^{-8}$

error = 0.118

error $= 0.\underbrace{00}147$

error $= 0.\underbrace{000000}22$

$K=3, \quad a=5 \qquad \sqrt[k]{a} = \underline{\underline{\sqrt[3]{5}}}$

Example. Let $k \geq 2$, and $a > 0$. Use the Newton's method to find an iterative scheme to approximate $\sqrt[k]{a}$. Use the formula to estimate $\sqrt[3]{2}$ starting with $x_0 = 1.2$.

We set up an algebraic function so that $\underline{\underline{\sqrt[k]{a}}}$ is one of its roots.

$$f(x) = x^k - a$$

$$f'(x) = k x^{k-1}$$

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^k - a}{k x_n^{k-1}} = \frac{(k-1) x_n^k + a}{k x_n^{k-1}}$$

For $k=3$ and $a=2$, we get an iterative scheme to estimate $\sqrt[3]{2}$.

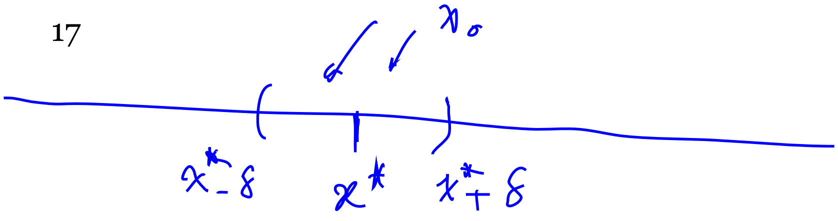
$$\begin{cases} x_0 = 1.2 \\ x_{n+1} = \dfrac{2x_n^3 + 2}{3x_n^2} \end{cases}$$

n	x_n	$\lvert\epsilon_a\rvert$
0	1.2	—
1	1.2630	0.0525
2	1.2599	0.0024
3	1.2599	0.0000058

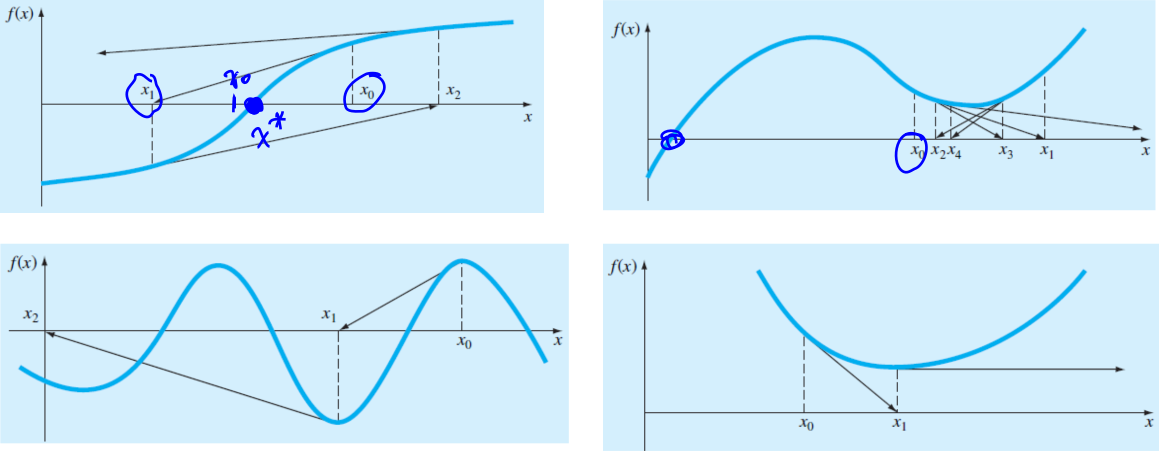
Local convergence analysis. Let $f(x)$ be a twice continuously differentiable function on $[a, b]$. If $x^* \in (a, b)$ is such that $f(x^*) = 0$ and $f'(x^*) \neq 0$, then there exists a $\delta > 0$ such that Newton's method generates a sequence $\{x_n\}$ converging to x^* for any initial approximation $x_0 \in [x^* - \delta, x^* + \delta]$. Moreover, the convergence rate is quadratic.

x_0

$\{x_n\}$



Note. The Newton method can perform very poor in some cases. See the following illustrative examples.



Example (try it yourself). Consider the van der Waals equation:

$$\left(P + \frac{n^2 a}{V^2}\right)(V - nb) = nRT$$

that relates the pressure (P), volume (V), and temperature (T) of a gas. Starting with $V_0 = 12.84$, use 3 iterations of Newton's method to find the volume of chlorine gas with $a = 6.29$ atm. liter^2/mole^2, $b = 0.0562$ liter/mole, $P = 2$ atmospheres, $n = 1$ mole, and $T = 313\ K$. , $R = 10$

$$f(V) = nRT - \left(p + \frac{n^2 a}{V^2}\right)(V - nb)$$

$$f'(V) = -\left(p + \frac{n^2 a}{V^2}\right) + \frac{2n^2 a}{V^3}(V - nb)$$

$$V_{n+1} = V_n - \frac{f(V_n)}{f'(V_n)}$$

n	V_n
0	12.84
1	1594.7
2	1565.1
3	1565.1

Note. The Newton's method is a fixed-point method when $g(x)$ is defined as below:

$$\boxed{g(x) = x - \frac{f(x)}{f'(x)}}$$

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

It is shown that the Newton's algorithm is quadratically convergent algorithm. It means that the rate of convergence is quadratic

$$e_{n+1} \leq M e_n^2$$

Secant method

Despite having a fast convergence rate in Newton's method, there are some problems that limit its deployment in practice. One of the problems is its local convergence property. Another one is the evaluation of the derivative. There are certain functions whose derivatives may be difficult or inconvenient to evaluate. For these cases, the derivative can be approximated by a backward finite divided difference:

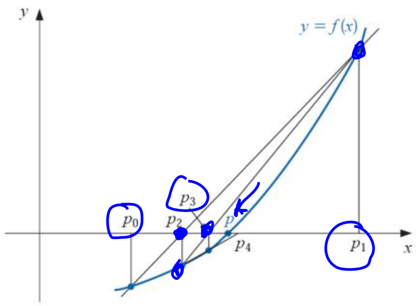
$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$$

$$f'(x_i) \cong \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}}$$

Thus, the iterative scheme for the secant method is:

x_0, x_1

$$x_{i+1} = x_i - \frac{f(x_i)(x_i - x_{i-1})}{f(x_i) - f(x_{i-1})}$$



Notice that the approach requires two initial estimates of x. Rather than using two arbitrary values to estimate the derivative, an alternative approach involves a fractional perturbation of the independent variable to estimate $f'(x)$, that leads to the following modified secant method:

x_0

$$x_{i+1} = x_i - \frac{\delta x_i f(x_i)}{f(x_i + \delta x_i) - f(x_i)}$$

where δ = a small perturbation fraction.

Example. Use the modified secant method to determine the mass of the bungee jumper using the equation $\sqrt{\frac{gm}{c_d}}\tanh\left(\sqrt{\frac{gc_d}{m}}t\right) - v = 0$ with $c_d = 0.25$ kg/m, $v = 36$ m/s after $t = 4$ s of free fall. The acceleration of gravity is 9.81 m/s^2. Use an initial guess of 50 kg and a value of 10^{-6} for the perturbation fraction.

Solution. 1st iteration:

$$x_0 = 50 \quad f(x_0) = -4.57938708$$

$$x_0 + \delta x_0 = 50.00005 \qquad f(x_0 + \delta x_0) = -4.579381118$$

0.434

$$x_1 = 50 - \frac{10^{-6}(50)(-4.57938708)}{-4.579381118 + 4.57938708} = 88.39931 \quad (|\varepsilon_t| = 38.1\%;\ |\varepsilon_a| = 43.4\%)$$

2nd iteration:

$$x_1 = 88.39931 \quad f(x_1) = -1.69220771$$

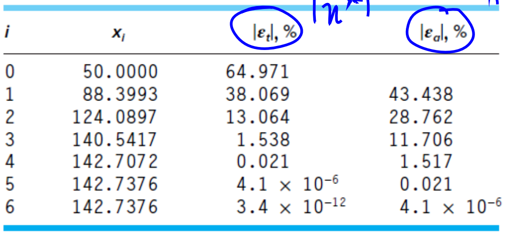
$$x_1 + \delta x_1 = 88.39940 \qquad f(x_1 + \delta x_1) = -1.692203516$$

$$x_2 = 88.39931 - \frac{10^{-6}(88.39931)(-1.69220771)}{-1.692203516 + 1.69220771} = 124.08970$$

$$(|\varepsilon_t| = 13.1\%;\ |\varepsilon_a| = 28.76\%) \rightarrow 0.2876$$

The calculation can be continued to yield: $\frac{|x_{n+1} - x^*|}{|x^*|}$

$\frac{|x_{n+1} - x_n|}{|x_{n+1}|}$



i	x_i	$\lvert\varepsilon_t\rvert$, %	$\lvert\varepsilon_a\rvert$, %
0	50.0000	64.971	
1	88.3993	38.069	43.438
2	124.0897	13.064	28.762
3	140.5417	1.538	11.706
4	142.7072	0.021	1.517
5	142.7376	4.1×10^{-6}	0.021
6	142.7376	3.4×10^{-12}	4.1×10^{-6}

Compute the third iteration (practice at home)

Example (try it yourself!) Use the first 7 iteration of the secant method with initial values $x_0 = 7$ and $x_1 = 8$ to estimate the root of $x^3 - \sinh x + 4x^2 + 6x + 9 = 0$.

Local convergence rate. Assume that x^* is the root of the twice differentiable function $f(x)$, that is $f(x^*) = 0$, and $f''(x^*) \neq 0$. Then, there exists $\delta > 0$ so that for any choices of $x_0, x_1 \in (x^* - \delta, x^* + \delta)$, the sequence $\{x_n\}$ generated by the secant method converges to x^*. Moreover, the rate of convergence is $p = \frac{1+\sqrt{5}}{2} \approx 1.62$.

$$e_{n+1} \leq M e_n^{1.62}$$

References

1. Chapra, Steven C. (2018). *Numerical Methods with* MATLAB *for Engineers and Scientists*, 4th Ed. McGraw Hill.
2. Bradie, Brian (2006), *A Friendly Introduction to Numerical Analysis*, 1st Ed., Pearson Prentice Hall.
3. Burden, Richard L., Faires, J. Douglas (2011). *Numerical Analysis*, 9th Ed. Brooks/Cole Cengage Learning