Chapter 5. Numerical Solution of Nonlinear Equations

1 Introduction to Iterative Methods

Finding one or more roots of an equation

$$f(x) \;=\; 0, \tag{1.1}$$

is one of the most commonly occurring problems in Applied Mathematics. Even the simplest of nonlinear equations – e.g., algebraic equations – are known to not admit solutions that are expressible rationally in terms of the data. It is therefore impossible, in general, to compute roots of nonlinear equations in a finite numbers of arithmetic operations.

The function $f : \mathbb{R}^m \to \mathbb{R}^n$ is a nonlinear function, which will be assumed to have a certain degree of smoothness. If $n > 1$, then (1.1) represents a system of n equations (at least one nonlinear) with m unknowns.

For now, we will restrict our discussion to the case $m = n = 1$, although many of the procedures we describe can easily be generalized to the multidimensional case.

Definition 1.1. *A number $\alpha \in \mathbb{C}$ satisfying equation* (1.1) *is called a **zero** or a **root** of f.*

As mentioned before, in most cases, explicit solutions of equation (1.1) are not available and we must try to find a root to any specified degree of accuracy. The numerical methods for finding the roots will be *iterative methods* and will require the knowledge of one (or more) initial value(s) x_0 ($x_1, \dots$). Then the method will produce a sequence $\{x_n\}_{n\in\mathbb{N}}$ of approximations of α, such that $\lim_{n\to\infty} x_n = \alpha$. These initial values will be determined, in general, from the context of the problem or from the graph of the function.

The analysis of an iterative method will include

– the proof of convergence, $x_n \to \alpha$, as $n \to \infty$;

– finding the *interval of convergence*, i.e. the set of values of the initial guess(es) x_0 ($x_1, \dots$) for which the method converges;

– determining the *speed* of convergence.

What makes an iterative method better than another is *how fast* it converges to the desired solution. Regarding the speed of convergence, we define the following:

Definition 1.2. *We say that a sequence of iterates $\{x_n\}_{n\in\mathbb{N}}$ converges to α with* ***order of convergence*** *$p \geq 1$, if*

$$|x_{n+1} - \alpha| \leq c\,|x_n - \alpha|^p, \text{ for all } n \in \mathbb{N}, \tag{1.2}$$

where $c > 0$ is a constant independent of n.
If $p = 1$, the method is said to ***converge linearly*** *to α, in which case we also require that $c < 1$. Then the constant c is called the* ***rate of linear convergence*** *of x_n to α.*
For $1 < p < 2$, we say that the convergence is ***superlinear****.*

Remark 1.3. If $p = 1$, then

$$|x_n - \alpha| \;\leq\; c\,|x_{n-1} - \alpha| \;\leq\; \ldots \;\leq\; c^n\,|x_0 - \alpha|,$$

which is why we require that $c < 1$.

2 Common Rootfinding Methods

We start with three simple methods and then give a general theory for one-point iteration methods. We recall some known results from Analysis, that will be used in the sequel.

Theorem 2.1. **[Intermediate Value Theorem]**
If $f : [a, b] \to \mathbb{R}$ is a continuous function, then it takes on any given value between $f(a)$ and $f(b)$ at some point within the interval. As a consequence, if a continuous function has values of opposite sign inside an interval $[a, b]$, then it has at least one root in that interval.

Theorem 2.2. **[Rolle's Theorem]**
If a function f is continuous on $[a, b]$ and differentiable on (a, b), with $f(a) = f(b)$, then there exists a point $c \in (a, b)$ such that $f'(c) = 0$. As a consequence, between any two distinct real roots of f, there is a root of the derivative.

So, combining the two, we can find the number of real zeros of a function (satisfying the conditions above) and locate them, by counting the number of *sign changes* of the function at the roots of the derivative and endpoints of the domain of definition.

2.1 Bisection Method

Assume that $f : \mathbb{R} \to \mathbb{R}$ is continuous on an interval $[a, b] \subset \mathbb{R}$ and that

$$f(a)f(b) \;<\; 0. \tag{2.1}$$

Then, by the Intermediate Value Theorem, there exists $\alpha \in (a, b)$ such that $f(\alpha) = 0$.

The simplest numerical procedure for finding a root is to repeatedly *halve (bisect)* the interval $[a, b]$, keeping the half on which $f(x)$ changes sign. This procedure is called the **bisection method**. Denoting by $[a_1, b_1] = [a, b]$, the method will produce a sequence of embedded intervals $[a_n, b_n]$, such that for every $n \in \mathbb{N}, \alpha \in [a_n, b_n], f(a_n)f(b_n) < 0$, and a sequence of approximations

$$c_n = \frac{a_n + b_n}{2} \tag{2.2}$$

of the root α (see Figure 1).

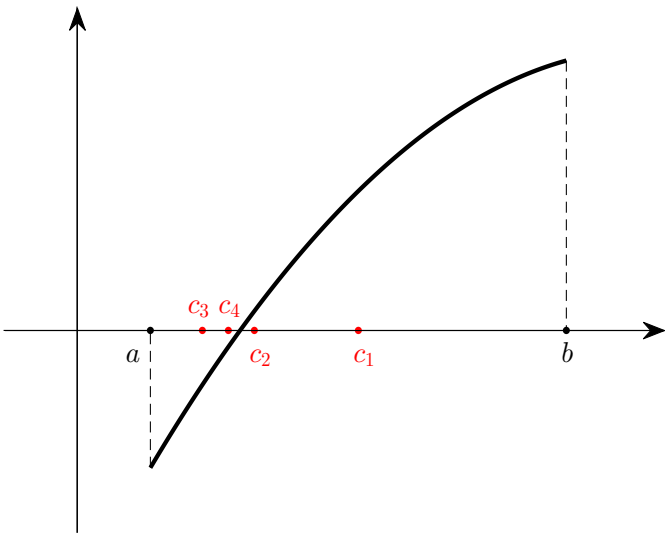

Fig. 1: Bisection method

Usually $[a, b]$ is chosen to contain only one root α, but the following algorithm for the bisection method will always converge to some root $\alpha \in [a, b]$, because of (2.1).

Algorithm 2.3. [Bisection method]
function $\alpha = \text{Bisect}(f, a, b, \varepsilon)$

1. Define $c = (a + b)/2$.

2. If $b - c \leq \varepsilon$, then $\alpha = c$ and exit.

3. If $\text{sign}(f(b)) \cdot \text{sign}(f(c)) \leq 0$, then $a = c$; otherwise, $b = c$.

4. Return to step 1.

The sequence $\{a_n\}_{n\in\mathbb{N}}$ is monotonely increasing, sequence $\{b_n\}_{n\in\mathbb{N}}$ is monotonely decreasing and

$$\lim_{n\to\infty} a_n \;=\; \lim_{n\to\infty} b_n \;=\; \lim_{n\to\infty} c_n \;=\; \alpha.$$

Also, we have

$$\begin{aligned} |x_n-\alpha| &\leq b_n - a_n = \frac{b-a}{2^n}, \\ |x_{n+1}-\alpha| &\leq \frac{1}{2}|x_n-\alpha|, \end{aligned} \tag{2.3}$$

which shows that the bisection method converges *linearly* (order of convergence $p=1$) with a rate of convergence of $\dfrac{1}{2}$.

Example 2.4. Find the largest root of

$$f(x) \;\equiv\; x^6 - x - 1 \;=\; 0, \tag{2.4}$$

with an error of $\varepsilon = 0.001$.

Solution. First, let us see how many real roots are there and where they are (approximately) located. We have

$$\begin{aligned} f(x) &= x^6 - x - 1, \\ f'(x) &= 6x^5 - 1. \end{aligned}$$

The derivative f' has only one real root, namely $\dfrac{1}{\sqrt[5]{6}}$. Now,

$$f\Big(\frac{1}{\sqrt[5]{6}}\Big) \;=\; \frac{1}{6}\cdot\frac{1}{\sqrt[5]{6}} - \frac{1}{\sqrt[5]{6}} - 1 \;=\; -\frac{5}{6}\cdot\frac{1}{\sqrt[5]{6}} - 1 \;<\; 0,$$

so the table of variation of f is

x	$-\infty$		$\dfrac{1}{\sqrt[5]{6}}$		∞
f	$+$		$-$		$+$

Thus, f has two real roots, one in $\Big(-\infty, \dfrac{1}{\sqrt[5]{6}}\Big)$ and one, $\alpha \in \Big(\dfrac{1}{\sqrt[5]{6}}, \infty\Big)$ (which we want to approximate).

In fact, since

$$\begin{aligned} f(-1) &= 1,\ f(0) = -1 \text{ and} \\ f(1) &= -1,\ f(2) = 61, \end{aligned}$$

we have a more precise location: a negative root between $(-1, 0)$ and the positive root that we seek, $\alpha \in (1, 2)$. That also gives us the starting interval for the bisection method. Alternatively, we can see from the graph the approximate location of the two real roots (see Figure 2).

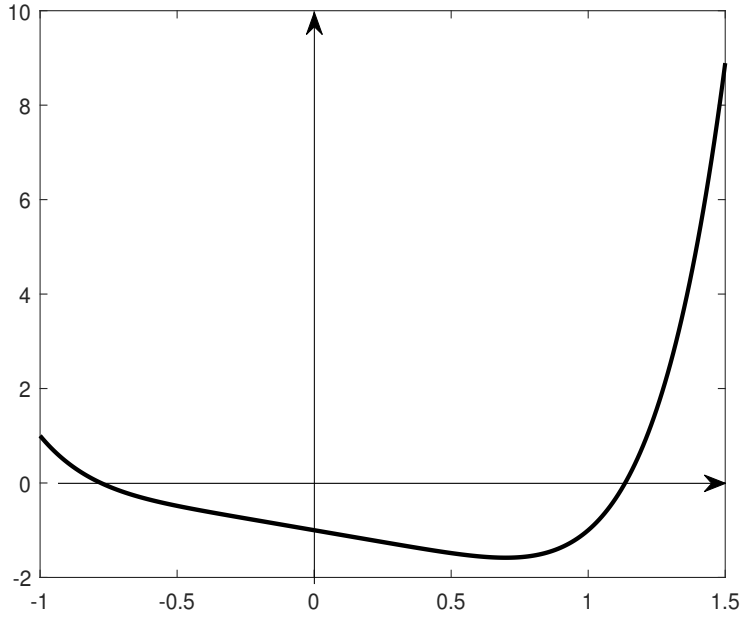

Fig. 2: Function $f(x) = x^6 - x - 1$

So, we start with the interval $[a_1, b_1] = [1, 2]$. How many iterations are needed for precision $\varepsilon = 0.001$? We find n from (2.3):

$$\begin{aligned} \frac{b-a}{2^n} &\leq \varepsilon, \text{ which means} \\ n &\geq \log_2\left(\frac{b-a}{\varepsilon}\right), \text{ i.e., in our example,} \\ n &\geq \log_2\left(\frac{1}{10^{-3}}\right) = 9.9658. \end{aligned}$$

The results of the bisection method are shown in Table 1. Indeed, after $n = 10$ iterations, we obtain the desired precision.

■

n	a_n	b_n	c_n	$b_n - c_n$	$f(c_n)$
1	1.0000	2.0000	1.5000	0.5000	8.8906
2	1.0000	1.5000	1.2500	0.2500	1.5647
3	1.0000	1.2500	1.1250	0.1250	−0.0977
4	1.1250	1.2500	1.1875	0.0625	0.6167
5	1.1250	1.1875	1.1562	0.0312	0.2333
6	1.1250	1.1562	1.1406	0.0156	0.0616
7	1.1250	1.1406	1.1328	0.0078	−0.0196
8	1.1328	1.1406	1.1367	0.0039	0.0206
9	1.1328	1.1367	1.1348	0.0020	0.0004
10	1.1328	1.1348	1.1338	0.00098	−0.0096

Table 1: Bisection Method for $x^6 - x - 1 = 0$

Remark 2.5. The bisection method is a *two-point method*, since two approximate values are needed to obtain an improved value. There are several advantages to the bisection method. The principal one is that the method is guaranteed to converge, as long as the function f is continuous and (2.1) is satisfied. In addition, the error bound given in (2.3) is guaranteed to decrease by one half with each iteration. This relation can also be used as a stopping criterion, as was done in the previous example. The principal disadvantage of the bisection method is that it generally converges slowly (only linearly), more slowly than most other methods. Also, it only approximates real roots.

The next two methods follow the same idea: approximate f by a linear interpolation polynomial and find the root of that polynomial. In other words, the graph of $y = f(x)$ is approximated by a straight line and the x-intercept of that line is approximating the root of f.

2.2 Secant Method

Assume that two initial guesses to α are known and denote them by x_0 and x_1. We approximate f by its Lagrange polynomial at the nodes x_0 and x_1. So the graph of $y = f(x)$ is approximated by the *secant* line determined by the points $(x_0, f(x_0))$ and $(x_1, f(x_1))$. The root α of f is then approximated by x_2, the x-intercept of the secant line. We hope x_2 will be an improved approximation of α. This is illustrated in Figure 3.

Let us find the value of x_2. The equation of the secant line is

$$y - f(x_1) = \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_1).$$

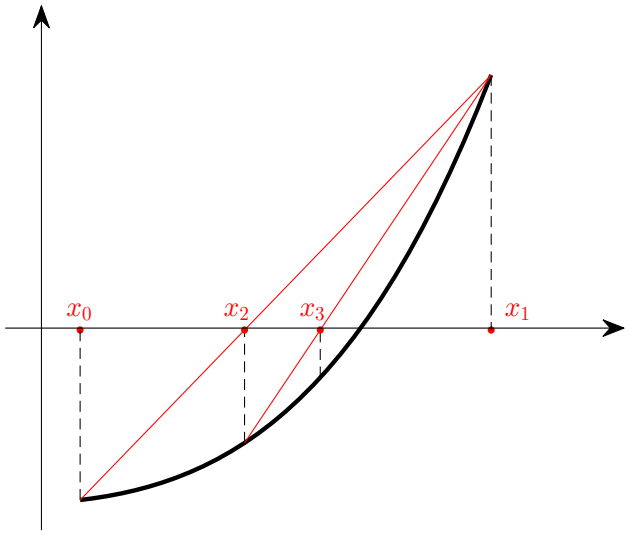

Fig. 3: Secant method

We find its point of intersection with the x-axis by letting $y = 0$ and solving for x. We get

$$x_2 \;=\; x_1 - f(x_1)\,\frac{x_1 - x_0}{f(x_1) - f(x_0)}$$

Having found x_2, we use x_1 and x_2 as a new set of approximate values for α. This leads to an improved value x_3. Recursively, we obtain a sequence of iterates given by

$$x_{n+1} \;=\; x_n - f(x_n)\,\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})},\; n = 1, 2, \ldots, \tag{2.5}$$

called the **secant method**.

Example 2.6. We solve again the equation

$$f(x) \;\equiv\; x^6 - x - 1 \;=\; 0,$$

which was used previously as an example for the bisection method.

Solution. We start with

$$x_0 \;= 1,\;\; x_1 \;=\; 2.$$

The results are given in Table 2, including the quantities $x_n - x_{n-1}$ as an estimate of $\alpha - x_{n-1}$. The

iterate x_8 equals α rounded to nine significant digits.

n	x_n	$f(x_n)$	$x_n - x_{n-1}$	$\alpha - x_{n-1}$
0	2.0	61.0		
1	1.0	-1.0	-1.0	
2	1.01612903	$-9.15e-1$	$1.61e-2$	$1.35e-1$
3	1.19057777	$6.57e-1$	$1.74e-1$	$1.19e-1$
4	1.11765583	$-1.68e-1$	$-7.29e-2$	$-5.59e-2$
5	1.13253155	$-2.24e-2$	$1.49e-2$	$1.71e-2$
6	1.13481681	$9.54e-4$	$2.29e-3$	$2.19e-3$
7	1.13472365	$-5.07e-6$	$-9.32e-5$	$-9.27e-5$
8	1.13472414	$-1.13e-9$	$4.92e-7$	$4.92e-7$

Table 2: Secant Method for $x^6 - x - 1 = 0$

■

The secant method is also a two-point iterative method. Unlike the bisection method, it *does not* always converge. For a convergence and error analysis, let us compute, from (2.5),

$$\begin{aligned} x_{n+1} - \alpha &= x_n - \alpha - f(x_n)\,\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} \\ &= x_n - \alpha - \frac{f(x_n)}{\dfrac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}} \\ &= x_n - \alpha - \frac{f(x_n) - f(\alpha)}{\dfrac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}}, \quad \text{since } f(\alpha) = 0. \end{aligned}$$

Further, we make use of divided differences and obtain

$$\begin{aligned} x_{n+1} - \alpha &= x_n - \alpha - (x_n - \alpha)\,\frac{f[x_n, \alpha]}{f[x_{n-1}, x_n]} \\ &= (x_n - \alpha)\left[1 - \frac{f[x_n, \alpha]}{f[x_{n-1}, x_n]}\right] \\ &= (x_n - \alpha)\,\frac{f[x_{n-1}, x_n] - f[x_n, \alpha]}{f[x_{n-1}, x_n]} \\ &= (x_n - \alpha)(x_{n-1} - \alpha)\,\frac{f[x_{n-1}, x_n, \alpha]}{f[x_{n-1}, x_n]}, \end{aligned}$$

so, assuming f is smooth enough,

$$x_{n+1} - \alpha \;=\; (x_n - \alpha)(x_{n-1} - \alpha)\,\frac{f''(\xi_n)}{2f'(\zeta_n)}, \tag{2.6}$$

with ζ_n between x_n and x_{n-1}, and ξ_n between the smallest and the largest of the numbers α, x_n and x_{n-1}. Using (2.6) and a limiting argument, we have the following convergence result.

Theorem 2.7. *Assume f, f' and f'' are continuous on an interval $I_\varepsilon = (\alpha - \varepsilon, \alpha + \varepsilon)$ containing the simple root α ($f'(\alpha) \neq 0$). Then, for starting values x_0 and x_1 sufficiently close to α, the iterates in* (2.5) *converge to α, with order of convergence*

$$p \;=\; r \;=\; \frac{1+\sqrt{5}}{2} \approx 1.618033\ldots, \tag{2.7}$$

*known as the **golden ratio**.*

Thus, the secant method converges *superlinearly*.

Remark 2.8.

1. To understand what "sufficiently close" means in the theorem above, let

$$M_\varepsilon \;=\; \frac{\max\limits_{I_\varepsilon} |f''(x)|}{2\min\limits_{I_\varepsilon} |f'(x)|},\; e_0 = |x_0 - \alpha|,\;\; e_1 = |x_1 - \alpha|. \tag{2.8}$$

Then the method above will converge if $x_0, x_1 \in I_\varepsilon$, with $\varepsilon > 0$ chosen so that

$$\max\{M_\varepsilon e_0, M_\varepsilon e_1\} < 1. \tag{2.9}$$

2. It is clear now that the secant method does not always converge, but when it does, it does so *faster* than the bisection method (its order of convergence is higher). That was obvious in our example.

3. Another advantage is that the secant method can be used to approximate complex roots, as well, if the initial values x_0 and x_1 are taken to be complex numbers satisfying the conditions above.

4. It can be shown that

$$\begin{aligned} \lim_{n\to\infty} \frac{|x_{n+1} - x_n|}{|x_n - \alpha|} \;&=\; 1 \text{ and, thus,} \\ |x_n - \alpha| \;&\approx\; |x_{n+1} - x_n|, \text{ for sufficiently large } n, \end{aligned} \tag{2.10}$$

which can be used as a stopping criterion.

2.3 Newton's Method

In a similar fashion, now we start with one initial value, x_0 and approximate f by its linear Taylor polynomial at the double node x_0. In other words, the graph of $y = f(x)$ is approximated by the line *tangent* to the graph of f at the point $(x_0, f(x_0))$. The root α of f is then approximated by x_1, the point of intersection of the tangent line with the x-axis. If x_0 is close enough to α, then the root of the Taylor polynomial should be close to α.

The tangent line at x_0 has equation

$$y - f(x_0) \quad = \quad f'(x_0)(x - x_0),$$

so, for x_1, we find

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

Repeat the process to further improve the estimate of α. Recursively, we get

$$x_{n+1} \quad = \quad x_n - \frac{f(x_n)}{f'(x_n)}, \; n = 0, 1, \ldots. \tag{2.11}$$

This is called **Newton's (tangent) method** and it is illustrated in Figure 4.

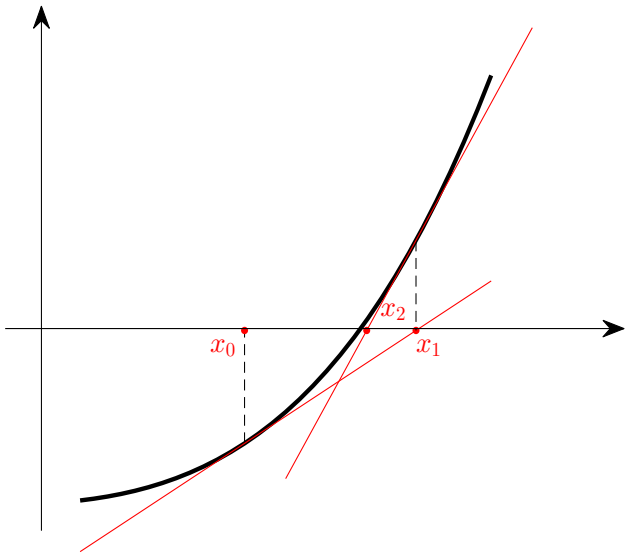

Fig. 4: Newton's method

Example 2.9. Let us approximate the positive solution of

$$f(x) \equiv x^6 - x - 1 = 0,$$

using Newton's method.

Solution. An initial guess x_0 can be taken from the graph of $y = f(x)$ in Figure 2. The iterative method is given by

$$x_{n+1} = x_n - \frac{x_n^6 - x_n - 1}{6x_n^5 - 1},\ n \geq 0.$$

Table 3 shows the results of Newton's method with initial value $x_0 = 1.5$.

n	x_n	$f(x_n)$	$x_n - x_{n-1}$	$\alpha - x_{n-1}$
0	1.5	$8.89e+1$		
1	1.30049088	$2.54e+1$	$-2.00e-1$	$-3.65e-1$
2	1.18148042	$5.38e-1$	$-1.19e-1$	$-1.66e-1$
3	1.13945559	$4.92e-2$	$-4.20e-2$	$-4.68e-2$
4	1.13477763	$5.50e-4$	$-4.68e-3$	$-4.73e-3$
5	1.13472415	$7.11e-8$	$-5.35e-5$	$-5.35e-5$
6	1.13472414	$1.55e-15$	$-6.91e-9$	$-6.91e-9$

Table 3: Newton's Method for $x^6 - x - 1 = 0$

As seen from the table, the convergence is very rapid. The iterate x_6 is accurate (almost) to the machine precision of around 16 decimal digits.

■

As before, we can compute

$$\begin{aligned} x_{n+1} - \alpha &= (x_n - \alpha)^2 \frac{f[x_n, x_n, \alpha]}{f[x_n, x_n]} \\ &= (x_n - \alpha)^2 \frac{f''(\xi_n)}{2f'(x_n)}, \end{aligned} \tag{2.12}$$

with ξ_n between α and x_n. Then we have the following convergence result.

Theorem 2.10. *Assume f, f' and f'' are continuous on an interval $I_\varepsilon = (\alpha - \varepsilon, \alpha + \varepsilon)$ containing the simple root α ($f'(\alpha) \neq 0$). Then, if the initial value x_0 is sufficiently close to α, the iterates in*

(2.11) *converge to* α *and*

$$\lim_{n\to\infty} \frac{x_{n+1}-\alpha}{(x_n-\alpha)^2} = \frac{f''(\alpha)}{2f'(\alpha)}, \tag{2.13}$$

which shows that the order of convergence of Newton's method is $p = 2$.

Remark 2.11.

1. Similarly with Remark 2.8, "sufficiently close" means $x_0 \in I_\varepsilon$, where ε is chosen so that $M_\varepsilon e_0 < 1$, with M_ε and e_0 defined in (2.8).

2. Again, as before, Newton's method *does not* always converge, but when it does, it does so faster ($p = 2$) than the bisection method ($p = 1$) and the secant method ($p = (1+\sqrt{5})/2 \approx 1.618$).

3. Also, Newton's method can be used to approximate complex roots, as well, if the initial value x_0 is a complex number satisfying the conditions above.

4. Again, for sufficiently large n,

$$|x_n - \alpha| \;\approx\; |x_{n+1} - x_n|,$$

which can be used as a stopping criterion.

5. Unlike the bisection and secant methods, Newton's method is a *one-step* iterative method, as it only requires one initial value. Later on, we will give a more comprehensive analysis of one-step iterative methods.

2.4 Comparison Between Newton's and Secant Methods

As we have seen, Newton's method and the secant method are closely related. If the approximation

$$f'(x_n) \;\approx\; \frac{f(x_n)-f(x_{n-1})}{x_n - x_{n-1}}$$

is used in Newton's formula (2.11), we obtain the secant formula (2.5).

The conditions for convergence are almost identical and the error formulas are similar. Nonetheless, there are two major differences. Newton's method requires two function evaluations per iterate, those of $f(x_n)$ and $f'(x_n)$, whereas the secant method requires only one function evaluation per iterate, that of $f(x_n)$ (provided that the value of $f(x_{n-1})$ is retained from the last iteration). So, Newton's method is generally more expensive per iteration. On the other hand, it converges more rapidly (order $p = 2$ versus $p = r \approx 1.62$) and consequently, it will require fewer iterations to attain a given desired accuracy. A comparison of the expenditure of computational time needed to approximate a root α within a desired tolerance, can be made.

To simplify the analysis, we assume that the initial guesses are quite close to the desired root, so both methods converge. Let t be the time needed to evaluate $f(x)$, and $s \cdot t$ the time required to evaluate $f'(x)$. By writing the operations involved in the two methods, it can be then shown that the minimum time to obtain the desired accuracy with Newton's method is

$$T_N \;=\; \frac{(1+s)tK}{\log 2},$$

while, for the secant method, a similar calculation shows that the minimum time necessary to obtain the desired accuracy is

$$T_S \;=\; \frac{tK}{\log r},$$

where K is a positive constant that depends on ε, x_0 and $c = \left|\frac{f''(\alpha)}{2f'(\alpha)}\right|$. Thus,

$$\frac{T_S}{T_N} \;=\; \frac{\log 2}{(1+s)\log r}.$$

The secant method is faster than Newton's method if the ratio is less than one, i.e.

$$\begin{aligned} \frac{T_S}{T_N} &< 1 \\ s &> \frac{\log 2}{\log r} - 1 \;\approx\; 0.44. \end{aligned}$$

In conclusion, if the time needed to evaluate $f'(x)$ is more than 44% of that necessary to evaluate $f(x)$, then the secant method is more efficient. In practice, many other factors will affect the relative costs of the two methods, so that the $.44$ factor should be used with caution.

3 One-Point Iteration Methods – General Theory

3.1 Fixed Point Iteration

A classical approach is to reformulate equation $f(x) = 0$ as

$$x \;=\; g(x) \tag{3.1}$$

and find a *fixed point* for g. Let us first note that the form (3.1) is *not* restrictive in any way. In fact, any equation can be written in the form (3.1) in a multitude of ways.

Example 3.1. Consider the equation

$$x^2 - 3 \;=\; 0.$$

It can be rewritten, for instance, as

$$\begin{aligned}
&\textbf{(a)} \quad x \;=\; x^2 + x - 3,\\
&\textbf{(b)} \quad x \;=\; \frac{3}{x},\\
&\textbf{(c)} \quad x \;=\; \frac{1}{2}\Big(x + \frac{3}{x}\Big),\\
&\textbf{(d)} \quad x \;=\; x + c(x^2 - 3), \text{ for some constant } c \in \mathbb{R},
\end{aligned}$$

and many other ways.

Now we can employ fixed point theory to discuss the solvability of equation (3.1). In what follows, the notation

$$g\big([a,b]\big) \;\subseteq\; [a,b]$$

means

$$x \in [a,b] \implies g(x) \in [a,b].$$

Lemma 3.2. *Let $g \in C[a,b]$, such that $g\big([a,b]\big) \subseteq [a,b]$. Then g has at least one fixed point in $[a,b]$.*

Proof. This follows immediately from the Intermediate Value Theorem applied to the function

$$G(x) = g(x) - x.$$

Since G is continuous and $G(a) \;\geq\; 0,\;\; G(b) \;\leq\; 0$, G must have at least one zero in $[a,b]$, which is, obviously, a fixed point of g.

□

Theorem 3.3. **[Banach]** *Let $g \in C[a,b]$, with $g\big([a,b]\big) \subseteq [a,b]$. Assume that there exists $0 < \lambda < 1$ such that*

$$\big|g(x) - g(y)\big| \;\leq\; \lambda\,|x - y|,\; \forall x, y \in [a,b] \;\text{ (i.e., } g \text{ is a } \textbf{contraction}\text{)}. \tag{3.2}$$

Then g has a unique fixed point $\alpha \in [a,b]$. Furthermore, the iterates

$$x_{n+1} \;=\; g(x_n),\ n \geq 0, \tag{3.3}$$

converge to α, for any choice of $x_0 \in [a,b]$ and the following error estimates hold:

$$\begin{aligned} |x_n - \alpha| &\leq \lambda\, |x_{n-1} - \alpha|,\ n \geq 1, \\ |x_n - \alpha| &\leq \frac{\lambda^n}{1-\lambda}\, |x_1 - x_0|. \end{aligned} \tag{3.4}$$

Proof. The existence of the fixed point is guaranteed by Lemma 3.2.
To prove its uniqueness, assume there are two fixed points, $\alpha = g(\alpha),\ \beta = g(\beta),\ \alpha \neq \beta$. Then

$$|\alpha - \beta| \;=\; \big|g(\alpha) - g(\beta)\big| \overset{(3.2)}{\leq} \lambda\, |\alpha - \beta|$$

and, so,

$$(1-\lambda)|\alpha - \beta| \;\leq 0,$$

which is a contradiction. Thus, $\alpha = \beta$.
To prove the convergence, let us note that

$$x_0 \in [a,b] \Longrightarrow x_1 = g(x_0) \in [a,b] \Longrightarrow \cdots \Longrightarrow x_n \in [a,b],\ \forall n \geq 0.$$

Then,

$$\begin{aligned} |x_n - \alpha| &= \big|g(x_{n-1}) - g(\alpha)\big| \leq \lambda \big|x_{n-1} - \alpha\big| = \lambda \big|g(x_{n-2}) - g(\alpha)\big| \\ &\leq \lambda^2 \big|x_{n-2} - \alpha\big| \leq \ldots \leq \lambda^n\, |x_0 - \alpha|. \end{aligned}$$

Letting $n \to \infty$, since $\lambda^n \to 0$, it follows that $x_n \to \alpha$ and the first bound in (3.4) holds.
For the second bound, we write

$$\begin{aligned} |x_0 - \alpha| &\leq |x_0 - x_1| + |x_1 - \alpha| \leq |x_0 - x_1| + \lambda\, |x_0 - \alpha|, \\ |x_0 - \alpha| &\leq \frac{1}{1-\lambda} |x_1 - x_0|. \end{aligned}$$

Combining this with the previous relation, we get the second bound in (3.4). □

Remark 3.4.
1. The first bound shows that $\{x_n\}_{n\in\mathbb{N}}$ converges *linearly*, with a rate of convergence bounded by

the contraction constant λ.

2. From the proof of Theorem 3.3, we can also show that

$$\begin{aligned} |x_n - \alpha| &\leq \frac{1}{1-\lambda}|x_{n+1} - x_n| \text{ and, hence,} \\ |x_{n+1} - \alpha| &\leq \lambda\,|x_n - \alpha| \leq \frac{\lambda}{1-\lambda}|x_{n+1} - x_n|, \end{aligned}$$

which gives a stopping criterion

$$|x_{n+1} - x_n| \leq \frac{1-\lambda}{\lambda}\,\varepsilon. \tag{3.5}$$

3. If g is also differentiable on (a, b), then, by the MVT, there exists $c \in (a, b)$ such that

$$g(x) - g(y) = g'(c)(x - y), \forall x, y \in [a, b].$$

Letting $\lambda = \max\limits_{x\in[a,b]} \big|g'(x)\big|$, it follows that

$$\big|g(x) - g(y)\big| \leq \lambda\,|x - y|,\ \forall x, y \in [a, b].$$

Then, we can restate the convergence result.

Theorem 3.5. *Let $g \in C^1[a, b]$, such that $g\big([a, b]\big) \subseteq [a, b]$ and*

$$\lambda := \max_{x\in[a,b]} \big|g'(x)\big| < 1. \tag{3.6}$$

Then:

***a)** Function g has a unique fixed point $\alpha \in [a, b]$.*

***b)** For any initial choice $x_0 \in [a, b]$, the sequence $x_{n+1} = g(x_n)$ converges to α, as $n \to \infty$.*

***c)** $|x_n - \alpha| \leq \lambda^n |x_0 - \alpha| \leq \dfrac{\lambda^n}{1-\lambda}|x_1 - x_0|,\ n \geq 1.$*

d)

$$\lim_{n\to\infty} \frac{x_{n+1} - \alpha}{x_n - \alpha} = g'(\alpha), \tag{3.7}$$

so, if $g'(\alpha) \neq 0$, the iterative method $x_{n+1} = g(x_n)$ is linearly convergent to the root α with rate of convergence bounded by λ.

The conditions of Theorem 3.5 can be relaxed and imposed only *locally*, near the root α.

Theorem 3.6. *Assume α is a fixed point of g and that g is continuously differentiable in a neighborhood of α, with*

$$|g'(\alpha)| \ < \ 1. \tag{3.8}$$

Then the conclusions of Theorem 3.5 still hold, provided that x_0 is chosen sufficiently close to α.

Example 3.7. Refer back to the equation $x^2 - 3 \ = \ 0$ in Example 3.1, with $\alpha = \sqrt{3}$. Let us see which of the four iterative methods are convergent, for x_0 sufficiently close to α.

Solution.
(a)

$$g(x) \ = \ x^2 + x - 3, \ g'(x) \ = \ 2x + 1, \ g'(\alpha) \ = \ 2\sqrt{3} + 1 \ > \ 1,$$

so this method *does not converge*.
(b)

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

so this method does not converge, either.
(c)

$$g(x) \ = \ \frac{1}{2}\Big(x + \frac{3}{x}\Big), \ g'(x) \ = \ \frac{1}{2}\Big(1 - \frac{3}{x^2}\Big), \ g'(\alpha) \ = \ 0.$$

This method *will converge* at least linearly.
(d)

$$g(x) \ = \ x + c(x^2 - 3), \ g'(x) \ = \ 1 + 2cx, \ g'(\alpha) \ = \ 1 + 2c\sqrt{3}.$$

For convergence, pick c such that $|g'(\alpha)| \ < \ 1$, i.e., $-\dfrac{1}{\sqrt{3}} \ < \ c \ < \ 0$.

For a good rate of linear convergence, pick c such that $1 + 2c\sqrt{3} \ \approx \ 0$, or $c \ \approx \ -\dfrac{1}{2\sqrt{3}}$, for example, $c \ = \ -\dfrac{1}{4}$.

■

Example 3.8. How many real roots does the equation

$$x - 1 - \arctan x \;=\; 0 \tag{3.9}$$

have? Will the iterative method

$$x_{n+1} \;=\; 1 + \arctan x_n \tag{3.10}$$

converge? For what starting values x_0? Find a bound for the error.

Solution. First, let us recall that the function $\arctan x$ is defined on the entire $\mathbb{R}$, but takes values *only* in the interval $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ (see its graph below). That means that

$$-\frac{\pi}{2} \;<\; \arctan x \;<\; \frac{\pi}{2}, \;\; \forall x \in \mathbb{R}.$$

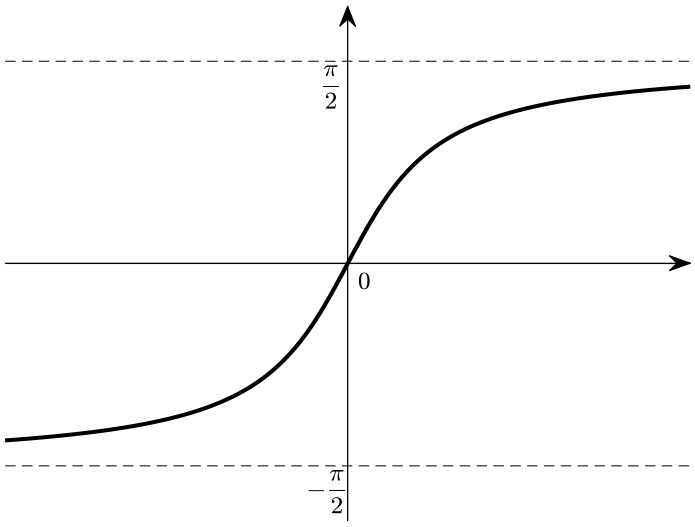

Fig. 5: Function $\arctan x$

Now, to find the number of real roots, let

$$f(x) \;=\; x - 1 - \arctan x.$$

Then

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

and is 0 only for $x = 0$. The table of variation of f is

x	$-\infty$	0	∞
f	$-$	$-$	$+$

So there is only one real root $\alpha > 0$. To locate it better, compute a few more values:

$$\begin{aligned} f(1) \;&=\; 1 - 1 - \frac{\pi}{4} \;=\; -\frac{\pi}{4} \;<\; 0, \\ f\Big(1+\frac{\pi}{2}\Big) \;&=\; 1 + \frac{\pi}{2} - 1 - \arctan\Big(1+\frac{\pi}{2}\Big) \;=\; \frac{\pi}{2} - \arctan\Big(1+\frac{\pi}{2}\Big) \;>\; 0, \end{aligned}$$

so $\alpha \in \Big(1, 1+\frac{\pi}{2}\Big)$.

To study the iterative method (3.10), let

$$g(x) \;=\; 1 + \arctan x.$$

Now equation (3.9) can be written in the fixed-point form $x \;=\; g(x)$ and the iteration (3.10) is given by $x_{n+1} \;=\; g(x_n)$.

Let us see if we can use Theorem 3.5, a *global* result, i.e., find an interval $[a, b]$ such that $g\big([a,b]\big) \;\subseteq\; [a,b]$. Since $\arctan x \;\leq\; \frac{\pi}{2}$, it follows that $g(x) \;=\; 1 + \arctan x \;\leq\; 1 + \frac{\pi}{2}$, for all $x \in \mathbb{R}$. Also,

$$g'(x) \;=\; \frac{1}{1+x^2},$$

which is strictly positive for all $x \;\in\; \mathbb{R}$. That means that g is strictly increasing on $\mathbb{R}$. So, for $x \in \Big[1, 1+\frac{\pi}{2}\Big]$, we have

$$g(1) \;\leq\; g(x) \;\leq\; g\Big(1+\frac{\pi}{2}\Big).$$

But

$$\begin{aligned} g(1) \;&=\; 1 + \frac{\pi}{4} \;>\; 1 \quad \text{and} \\ g\Big(1+\frac{\pi}{2}\Big) \;&=\; 1 + \arctan\Big(1+\frac{\pi}{2}\Big) \;<\; 1 + \frac{\pi}{2}. \end{aligned}$$

Thus,

$$g\left(\left[1, 1+\frac{\pi}{2}\right]\right) \subseteq \left[1, 1+\frac{\pi}{2}\right].$$

Now, $g''(x) = -\dfrac{2x}{(1+x^2)^2}$, which is strictly negative on $\left[1, 1+\dfrac{\pi}{2}\right]$, so g' is strictly decreasing on that interval. Then, for all $x \in \left[1, 1+\dfrac{\pi}{2}\right]$,

$$g'(x) \leq g'(1) = \frac{1}{2} < 1.$$

So, by Theorem 3.5 with $\lambda = \dfrac{1}{2}$, the iteration (3.10), $x_{n+1} = g(x_n) = 1 + \arctan x_n$ converges to α, for any starting value $x_0 \in \left[1, 1+\dfrac{\pi}{2}\right]$ and we have the error estimate

$$|x_n - \alpha| \leq \frac{1}{2^{n-1}} |x_1 - x_0|.$$

The exact solution with 10 correct decimals is $\alpha = 2.1322679602$. Indeed, the convergence of the iteration (3.10) is quite fast, as seen in Table 4, for various values of $x_0 \in \left[1, 1+\dfrac{\pi}{2}\right]$.

	$x_0 = 1$		$x_0 = 1+\pi/4$		$x_0 = 1+\pi/2$	
n	x_n	$\lvert x_n - \alpha\rvert$	x_n	$\lvert x_n - \alpha\rvert$	x_n	$\lvert x_n - \alpha\rvert$
1	1.78540	$3.47e-1$	2.06023	$7.20e-2$	2.19982	$6.76e-2$
2	2.06023	$7.20e-2$	2.11891	$1.34e-2$	2.14414	$1.19e-2$
3	2.11891	$1.34e-2$	2.12985	$2.42e-3$	2.13440	$2.13e-3$
4	2.12985	$2.42e-3$	2.13183	$4.37e-4$	2.13265	$3.84e-4$
5	2.13183	$4.37e-4$	2.13219	$7.90e-5$	2.13234	$6.89e-5$
6	2.13219	$7.90e-5$	2.13225	$1.44e-5$	2.13228	$1.22e-5$
7	2.13225	$1.44e-5$	2.13227	$2.80e-6$	2.13227	$2.01e-6$
8	2.13227	$2.80e-6$	2.13227	$6.97e-7$	2.13227	$1.70e-7$

Table 4: Example 3.8

■