

## 1. ITERATIVE SOLUTIONS OF NONLINEAR EQUATIONS

Before getting to our problem, we need some introductory notions.

**Definition.** - Let  $f : [a, b] \rightarrow \mathbf{R}$  and suppose that  $\bar{x} \in [a, b]$  is such that  $f(\bar{x}) = 0$ . We define order of the zero  $\bar{x}$  the least upper bound of all numbers  $k$  such that

$$\lim_{x \rightarrow \bar{x}} \frac{|f(x)|}{|x - \bar{x}|^k} < +\infty.$$

For example, both  $f_1(x) = |x|^{1/2}$  and  $f_2(x) = x^{1/2} \log x$  have in  $\bar{x} = 0$  a zero of order  $\frac{1}{2}$ .

We do not want to go into details here, but it is worth saying that the order of a zero is directly linked to the number of derivatives that vanish at  $\bar{x}$ .

When dealing with the numerical solution of mathematical problems, we have to take into account that there is a limit to the attainable accuracy. There are three sources of error:

- 1) It is usually necessary to replace  $f(x)$  with a simpler  $F(x)$ ; for example, if we need to work with  $f(x) = \sin x$ , then we might let  $F(x)$  be the polynomial corresponding to a given number of terms in the Taylor expansion for  $f$ .
- 2) The computation of  $F$  naturally involves rounding errors.
- 3)  $F(x)$  can only be evaluated for values of  $x$  which are computer representable numbers.

Let us finally come to our problem: we are interested in nonlinear equations, because for linear ones there is nothing to say.

**Problem.** - Find  $\alpha \in I = [a, b]$  s.t.

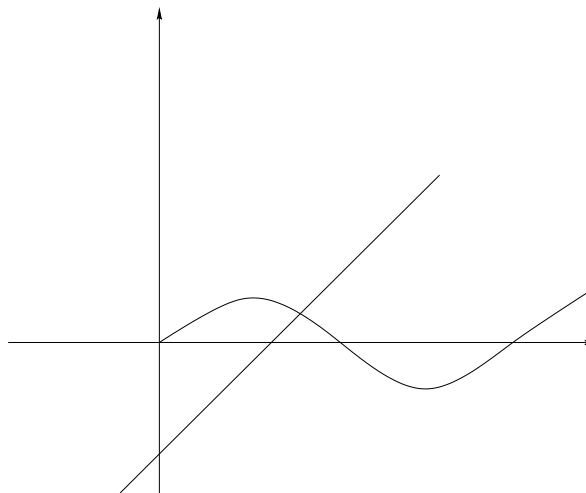
$$f(\alpha) = 0$$

where  $f$  is a given (possibly regular) one-variable function.

A first way to solve a nonlinear equation is the so-called *graphical method*; suppose for example that we need to solve

$$\sin x - x + 2 = 0 \quad \Leftrightarrow \quad \sin x = x - 2.$$

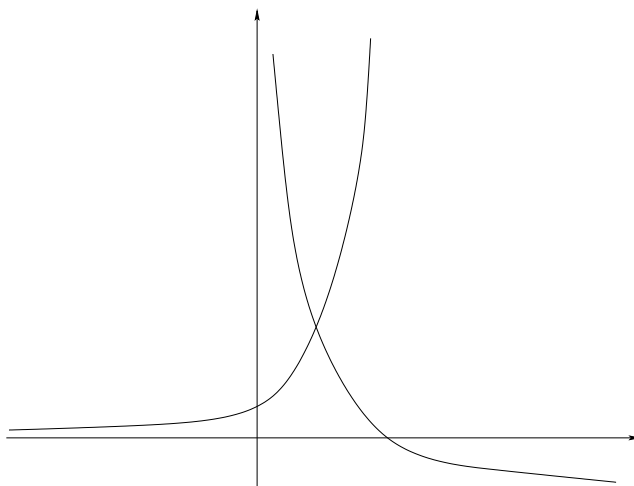
We can then set  $y_1 = \sin x$ ,  $y_2 = x - 2$ , draw the graphs both of  $y_1$  and  $y_2$  and in this way check directly if they have one or more intersection points.



In this case it is appasrent that they have a single intersection point. A similar situation occurs when solving

$$e^x + \ln x - 3 = 0 \quad \Leftrightarrow \quad e^x = 3 - \ln x,$$

as depicted below.



In this case we have set  $y_1 = e^x$ , and  $y_2 = 3 - \ln x$ .

It is rather obvious that this represents a sort of first aid method: it can help in understanding how things are, but in general it cannot provide a fine estimate of the actual value of the root we are looking for.

**Simple Idea.** - Use the Bisection Method

Let us recall the following result, which is the base for the bisection method:

**Intermediate Value Theorem.** - Let  $f : [a, b] \rightarrow \mathbf{R}$  be a continuous function such that  $f(a)f(b) < 0$ . Then there exists  $\alpha \in ]a, b[$  such that  $f(\alpha) = 0$ .

**Idea of the proof.** - Set  $a_o := a$ ,  $b_o := b$  and define  $x_o = \frac{a_o + b_o}{2}$ . Now if  $f(x_o) = 0$  we are finished. Otherwise

- a) if  $f(x_o) > 0$  set  $a_1 := x_o$ ,  $b_1 := b_o$ ;
- b) if  $f(x_o) < 0$  set  $a_1 := a_o$ ,  $b_1 := x_o$ ;
- c)  $x_1 = \frac{a_1 + b_1}{2}$ ,

and then iterate. We come up with a sequence of intervals  $I_k = [a_k, b_k]$  such that  $\forall k \geq 1$   $I_k \subset I_{k-1}$ . Moreover

$$|I_k| = \frac{|I|}{2^k} = \frac{b - a}{2^k}.$$

If we define the *absolute error*  $e_k = x_k - \alpha$  it is obvious that

$$|e_k| \leq |I_k| = \frac{b - a}{2^k} \rightarrow 0 \quad \text{as } k \rightarrow +\infty \quad \blacksquare.$$

Let us see the main features of this method.

**Positive Facts. -**

- a) It is globally convergent, i.e. there is no requirement on the starting step, except for the information on the value of  $f$  at the extreme points of the interval.
- b) We have a precise estimate on the speed of convergence. In fact if we want the error to be less than a fixed  $\epsilon$ , i.e.  $|x_k - \alpha| < \epsilon$ , all we need to do is to solve

$$\frac{b-a}{2^k} < \epsilon \quad \Rightarrow \quad k > \frac{\ln \frac{b-a}{\epsilon}}{\ln 2} = \frac{\ln \frac{b-a}{\epsilon}}{0.6931}.$$

Hence if we want

$$|x_k - \alpha| = \frac{|x_j - \alpha|}{10} \quad (\text{a gain of one decimal})$$

we must take

$$k - j = \log_2 10 \approx 3.32$$

that is, between three and four more iteration steps.

**Negative Facts. -**

- a) As it should be evident from the previous discussion, we have slow convergence of the method.
- b) It is easy to see that in general we do not have monotone reduction of the absolute error. Under this point of view, see Fig. 1 in the slides.

What we discussed above is a purely theoretical approach to the bisection method: if we come to real practice, then we need to define the accuracy we want to achieve. This is usually done prescribing two positive quantities  $\epsilon_1$  and  $\epsilon_2$ : we agree to accept  $\alpha$  as a root of our nonlinear equation, if either

$$|f(\alpha)| \leq \epsilon_1,$$

or if  $\alpha$  lies in the interval  $[\beta, \gamma]$  such that

$$f(\beta)f(\gamma) < 0, \quad \gamma - \beta \leq \epsilon_2.$$

Hence, given  $a$  and  $b$ , we first test whether  $|f(a)| \leq \epsilon_1$ , or  $|f(b)| \leq \epsilon_1$ : if so, we stop. We also test whether  $b - a \leq \epsilon_2$ : if so, we accept as root  $x_o = \frac{a+b}{2}$ . Otherwise, we start iterating.

If  $\epsilon_1$  and  $\epsilon_2$  are chosen with reasonable care, the above process will terminate, but it can continue indefinitely if  $\epsilon_1$  and  $\epsilon_2$  are too small.

This may happen if  $\epsilon_1$  is less than the expected rounding error for  $f$  in  $[a, b]$  and if  $\epsilon_2$  is less than the distance between two consecutive computer representable numbers in  $[a, b]$ .

The example presented in Fig. 1 in the slides suggests a two - step procedure:

- a) Apply the bisection method (slow but reliable) to get a first good but rough estimate of the root we are looking for.

- b) To the rough estimate apply a finer method which gives a better approximation with few further iterations.

**Natural Problem.** - How do I devise a finer method?

**First Idea.** - Suppose we write  $f(x) = x - g(x)$ . Then it is obvious that

$$f(\alpha) = 0 \quad \Leftrightarrow \quad \alpha = g(\alpha).$$

**Second Idea.** - By an iterative algorithm build the sequence

$$x_{k+1} = g(x_k).$$

Of course the hope is that  $x_k \rightarrow \alpha$ . Is it a sound hope? The answer is yes if ...

To give a satisfactory answer to the previous question, we rely on the following abstract result.

**Contractive Mapping Theorem.** - *Let  $X$  be a complete metric space endowed with the distance  $d$  and let  $g : X \rightarrow X$  be a mapping such that*

$$\exists L \in [0, 1[ \quad \text{s.t.} \quad \forall x_1, x_2 \in X \quad d(g(x_1), g(x_2)) \leq L d(x_1, x_2).$$

*Then there exists a unique  $\alpha \in X$  s.t.  $\alpha = g(\alpha)$ . Moreover for any starting point  $x_o \in X$  the sequence  $\{x_k = g(x_{k-1})\}$  converges to  $\alpha$ . Finally*

$$d(x_n, x_m) \leq \frac{L^m - L^n}{1 - L} d(x_1, x_o) \quad m < n,$$

$$d(\alpha, x_m) \leq \frac{L^m}{1 - L} d(x_1, x_o) \quad \blacksquare.$$

How can we apply it to our framework?

**Corollary.** - *Let  $g : [x_o - \rho_o, x_o + \rho_o] \rightarrow \mathbf{R}$  satisfy*

$$(*) \quad |g(x_1) - g(x_2)| \leq L|x_1 - x_2|$$

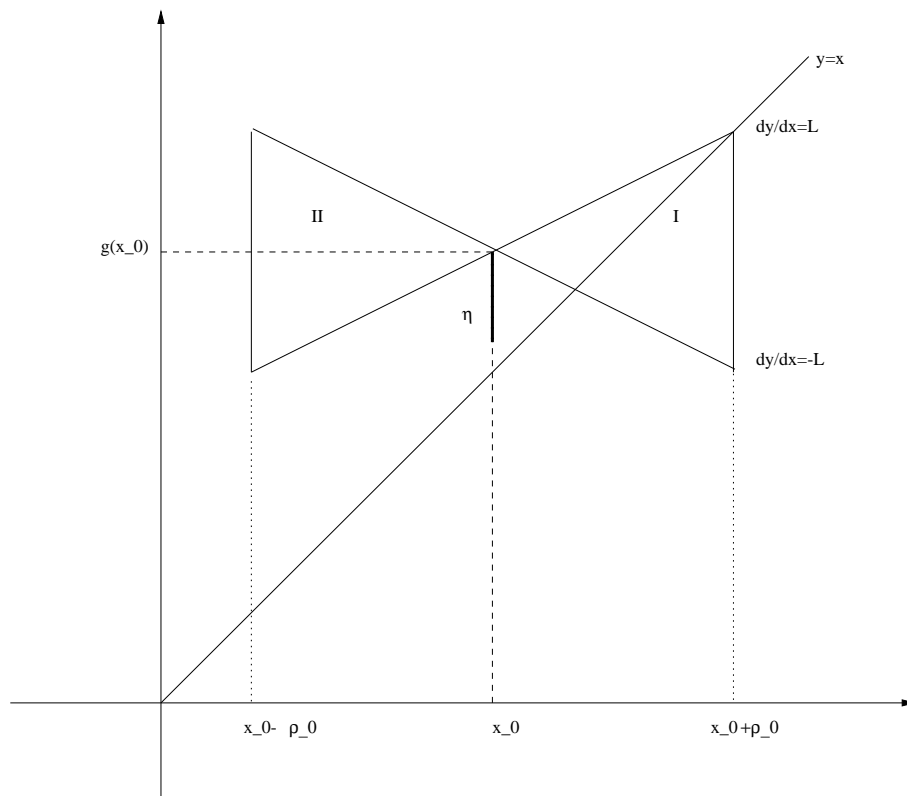
*for all values  $x_1, x_2$  in  $[x_o - \rho_o, x_o + \rho_o]$  with  $L \in [0, 1[$ . Let  $x_o$  be such that  $|x_o - g(x_o)| \leq (1 - L)\rho_o$ . Then*

- a)  $x_k \in [x_o - \rho_o, x_o + \rho_o]$ ;
- b)  $x_k \rightarrow \alpha$  and  $|x_k - \alpha| \leq L^k \rho_o$ ;
- c)  $\alpha$  is the only root in  $[x_o - \rho_o, x_o + \rho_o]$ .

Condition (\*) (a so - called Lipschitz condition) is usually hard to verify, but it is certainly satisfied if  $|g'(x)| \leq L$  for any  $x \in [x_o - \rho_o, x_o + \rho_o]$ . In this case we can also say that

$$\lim_{k \rightarrow +\infty} \frac{x_{k+1} - \alpha}{x_k - \alpha} = g'(\alpha).$$

Let us consider in the following figure the geometric interpretation of the previous Corollary.



$$\eta = g(x_0) - x_0$$

I and II are the regions in which the values of  $g(x)$  lie for  $x_0 - \rho_0 \leq x \leq x_0 + \rho_0$ . It is not hard to check that

- a) If  $L \geq 1$  the line  $y = x$  will not intersect the upper boundary of triangle I.
- b) If  $L < 1$  and  $\eta > (1 - L)\rho_0$ , again the line  $y = x$  will not intersect the upper boundary of triangle I.

The failed intersection implies that  $y = x$  may not intersect an admissible function  $g(x)$  in the interval  $[x_0 - \rho_0, x_0 + \rho_0]$  and the procedure would not converge.

In Fig. 2, you can see the general behaviour for  $g$  with positive slope and for  $g$  with negative slope.

Let us see the main features of the abstract method.

### Positive Facts. -

- a) Again it's a global method, because it converges for any  $x_0$  which satisfies the condition  $|x_0 - g(x_0)| < (1 - L)\rho_0$ .
- b) We have a precise estimate on the approximation error.

### Negative Facts. -

- a) How is  $\rho_0$  fixed?

- b) How do I check the Lipschitz condition? (which is somehow the same as asking who ensures that  $g$  is differentiable)
- c) How do I build  $g$  from  $f$  in practical situations?
- d) How do I control possible errors in the repeated evaluation of  $g$ ?

**Definition.** - When we have a sequence  $\{x_k\}$  which converges to  $\alpha$ , we say that the generating iteration scheme is a method of order  $p$  if

$$\exists C > 0 \text{ s.t. } \frac{|x_{k+1} - \alpha|}{|x_k - \alpha|^p} \leq C \quad \forall k \geq k_o$$

where  $k_o$  is a proper integer ■.

According to this definition, our abstract method is a first order method provided that  $g'(\alpha) \neq 0$ . In fact we have

$$|\alpha - x_k| = |g(\alpha) - g(x_{k-1})| \leq L|\alpha - x_{k-1}| \leq \dots \leq L^k|\alpha - x_o|.$$

The quantity  $\hat{L} := g'(\alpha)$  is called the *convergence factor* and  $R := \ln \frac{1}{g'(\alpha)}$  is called *rate of convergence*. This means that the number of additional iterations required to reduce the error at  $k$ -step by the factor  $10^{-m}$  is asymptotically  $\nu = \frac{m}{R}$ .

Suppose now that  $g'(\alpha) = 0$ . By Taylor's expansion we have

$$g(x) = g(\alpha) + g'(\alpha)(x - \alpha) + \frac{g''(\xi)}{2!}(x - \alpha)^2$$

$$g(x) = \alpha + \frac{g''(\xi)}{2!}(x - \alpha)^2.$$

If we repeat what we did before in this new context we get

$$|x_k - \alpha| = |g(x_{k-1}) - g(\alpha)| \leq \frac{1}{2}g''(\xi)|x_{k-1} - \alpha|^2.$$

Hence the error at any iterate is proportional to the square of the previous error and in this case we have a second order method.

Suppose now that  $|\frac{g''(x)}{2!}| \leq M$  in  $[\alpha - \rho_o, \alpha + \rho_o]$ . It is easy to see that

$$\begin{aligned} |x_k - \alpha| &\leq M|x_{k-1} - \alpha|^2 \\ &\leq M \cdot M^2|x_{k-2} - \alpha|^4 \\ &\vdots \\ &\leq (M|x_o - \alpha|)^{2^k - 1}|x_o - \alpha| \end{aligned}$$

and therefore if  $M|x_o - \alpha| < 1$  (remember what we said before about a good estimate to start with) the second order method reduces the initial error by a factor of at least  $10^{-m}$  when

$$(M|x_o - \alpha|)^{2^k - 1} \approx 10^{-m}.$$

The number  $\nu$  of required iterations is now

$$2^\nu - 1 \approx \frac{-m}{\ln(M|x_o - \alpha|)}.$$

If we assume in the previous first order method  $\hat{L} = M|x_o - \alpha|$ , we can compare the two methods and see that we have equal reduction of error if

$$2^{\nu_2} = 1 + \nu_1,$$

which means that 7 iterations for the second order method are approximately equivalent to 130 iterations for the first order one.

We could also show that the number of correct decimals more than double on each iteration of the second order method. Things are obviously better for higher order methods, but we won't consider them in the following.

**Natural Goal.** - Build a reliable (and possibly second or even higher order) method based on the previous abstract scheme.

The crucial point is the right choice of  $g$ , besides the simplest one we suggested at the beginning. Let us see two different possibilities.

- a) Let  $\Phi : [a, b] \rightarrow \mathbf{R}$  s.t.  $\forall x \in [a, b]$  we have  $0 < |\Phi(x)| < +\infty$ . Then we can set  $g(x) = x - \Phi(x)f(x)$ .
- b) Let  $F : \mathbf{R} \rightarrow \mathbf{R}$  s.t.  $F(0) = 0$  and  $\forall y \neq 0 F(y) \neq 0$ . Then we can set  $g(x) = x - F(f(x))$ .

The first choice is the standard one, the second one more naturally leads to higher order methods. In the following we will consider some natural choices for  $\Phi$  with the associated iterative methods.

## 1) CHORD METHOD

Set  $\Phi(x) = m$  with  $m$  constant. Then

$$g(x) = x - m f(x), \quad g'(x) = 1 - m f'(x).$$

As we need  $|g'(\alpha)| < 1$ , this requires  $-1 < 1 - m f'(\alpha) < 1$ , that is,  $0 < m f'(\alpha) < 2$ . Hence  $m$  must have the same sign as  $f'(\alpha)$ . Unfortunately if  $f'(\alpha) = 0$  the previous condition cannot be satisfied. Otherwise, if we assume

$$m = \frac{b - a}{f(b) - f(a)}$$

we obtain a bound on the interval to start with, namely

$$0 < \frac{b - a}{f(b) - f(a)} f'(\alpha) < 2 \quad \Rightarrow \quad b - a < 2 \frac{f(b) - f(a)}{f'(\alpha)}.$$

**General Features of the Chord Method.** -

- a) It's a first order method, except in the lucky case of  $f'(\alpha) = \frac{f(b)-f(a)}{b-a}$  which guarantees that  $g'(\alpha) = 0$ .
- b) It has slow convergence. It works well only when the starting interval  $[a, b]$  is finely tuned around  $\alpha$ .

The method has a natural geometric interpretation. In fact the sequence is built according to the rule

$$x_{k+1} = x_k - \frac{b-a}{f(b)-f(a)} f(x_k).$$

If we consider the line through  $(x_k, f(x_k))$  with slope  $\frac{1}{m}$ , then  $x_{k+1}$  is the  $x$ -intercept of such a line. So we can say that the next iterate is determined by a chord of constant slope joining a point on the curve to the  $x$ -axis (which justifies the name of the method!). You can see a pictorial description of this in Fig. 3.

## 2) NEWTON'S METHOD

With respect to the previous method, suppose that the slope of the chord is changed at each iteration so that

$$g'(x_k) = 1 - m_k f'(x_k) = 0$$

in order to obtain a second order method. From the previous relation, we obtain

$$m_k = \frac{1}{f'(x_k)}$$

which suggests

$$\Phi(x) = \frac{1}{f'(x)} \quad \Leftrightarrow \quad g(x) = x - \frac{f(x)}{f'(x)}.$$

We now have the sequence

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}.$$

Notice that  $g'(x) = \frac{f(x)f''(x)}{[f'(x)]^2}$ . Hence if  $f'(\alpha) \neq 0$  and  $f''(\alpha)$  exists, we come up with a second order method.

### General Features of Newton's Method. -

- a) It's a second order method (but see the discussion below in a particular case).
- b) It's fast.
- c) It's heavy under a computational point of view, because at each step requires evaluations both of  $f$  and  $f'$ .
- d) It may be impractical to evaluate  $f'$  at any step if  $f$  is known only implicitly.

Once more the method has a natural geometric interpretation. In fact the sequence is built according to the rule

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}.$$

If we consider the line through  $(x_k, f(x_k))$  which is tangent to the graph of  $f$  at that point (i.e. with slope  $m = f'(x_k)$ ), then  $x_{k+1}$  is the  $x$ -intercept of such a line. You can see a pictorial description of this in Fig. 4.

If  $f'(\alpha) = 0$  (i.e.  $\alpha$  is not a simple root of  $f(x) = 0$ ) then we are back to a first order method and  $g'(\alpha) = 1 - \frac{1}{p}$ , where  $p$  is the order of the root. If such a order is known, then we can fix things if we assume

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

A further remark on Newton's method is that it can be used also for complex valued functions. In that case, we have

$$z_{n+1} = z_n - \frac{f(z_n)}{f'(z_n)},$$

or equivalently

$$x_{n+1} = x_n - \frac{uu_x + vv_x}{u_x^2 + u_y^2} \Big|_{z_n},$$

$$y_{n+1} = y_n - \frac{vu_x - uv_x}{u_x^2 + u_y^2} \Big|_{z_n},$$

where we have set  $z = x + iy$ ,  $f(z) = f(x + iy) = f(x, y) = u(x, y) + iv(x, y)$ .

Notice that if  $f$  is real for real  $z$ , as it is the case for  $f(z) = z^2 + 1$ , in order to get complex roots, we must let the starting value  $z_o$  be complex, otherwise all iterants will be real. In fact for  $f(z) = z^2 + 1$  we have

$$z_{n+1} = \frac{z_n^2 - 1}{2z_n},$$

and if  $z_o$  is real, so will be  $z_1, z_2$ , etc.

### 3) SECANT METHOD

We saw before that the evaluation of  $f'(x_k)$  can be complicated or even impossible (at least with a sufficient degree of precision). The natural idea is then to substitute it with its discretized version

$$\frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}.$$

We now have

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} f(x_k).$$

Strictly speaking we are not dealing with an application of the Contractive Mapping Theorem. Moreover notice that to start the method we need two approximations of  $\alpha$ .

Even in this case the method has a natural geometric interpretation. If we consider the line through  $(x_k, f(x_k))$  and  $(x_{k-1}, f(x_{k-1}))$ , then  $x_{k+1}$  is the  $x$ -intercept of such a line. You can see a pictorial description of this in Fig. 5.

### General Features of the Secant Method. -

- a) It's fast
- b) It is lighter than Newton's Method under a computational point of view. It may look that we are still computing both  $f(x_k)$  and  $f(x_{k-1})$  at each iterations, but  $f(x_{k-1})$  may actually be retained from the previous step.
- c) It's a higher order method. Since we cannot directly write  $x_{k+1} = g(x_k)$ , the estimate of the order of the method cannot be done as before. Anyway with some calculations the order turns out to be  $p = \frac{1+\sqrt{5}}{2} \approx 1.618$ .

### 3) REGULA FALSI OR METHOD OF FALSE POSITION

With respect to the previous method, instead of  $(x_{k-1}, f(x_{k-1}))$ , we can use the latest point  $x_{k'}$  for which the function value  $f(x_{k'})$  has sign opposite that of  $f(x_k)$ , that is we have

$$x_{k+1} = x_k - \frac{x_k - x_{k'}}{f(x_k) - f(x_{k'})} f(x_k), \quad f(x_k) f(x_{k'}) < 0.$$

As before we need to start with two approximations of  $\alpha$  with the further requirement that the corresponding function values have opposite sign.

### General Features of the Method of False Position. -

- a) It's a first order method.
- b) It's generally faster than the bisection method, but slower than Newton's or the secant methods.
- c) The sequence of indices  $k'$  is non decreasing so that to determine  $k'$  at step  $k$  we don't need to check the whole sequence but just to stop at the value  $k'$  used at step  $k - 1$ .
- d) All the iterates are contained in the interval  $[x_{-1}, x_0]$  and this gives an upper bound on the error (this does not happen with the secant method).
- e) The sequence stops when  $|f(x_k)| < \epsilon$ ,  $\epsilon$  being a proper tolerance parameter fixed at the beginning.

In some peculiar instances, the convergence of the method of false position may be very slow. Consider the function

$$f(x) = \begin{cases} \delta, & \text{if } x \in [0, \frac{1}{2}], \\ 4(1 + \delta)(x - x^2) - 1, & \text{if } x \in (\frac{1}{2}, 1], \end{cases}$$

where  $\delta$  is a proper parameter in  $(0, 1)$ . It is easy to see that applying the rule of false position, we have

$$\begin{aligned} x_1 &= 1 - \frac{1}{1 + \delta}, \\ x_2 &= 1 - \frac{1}{(1 + \delta)^2}, \\ &\vdots \end{aligned}$$

$$x_n = 1 - \frac{1}{(1 + \delta)^n},$$

provided  $n$  is such that  $x_{n-1} \leq \frac{1}{2}$ . Suppose that  $\delta$  is very small. The number of iterations so that  $x_n \geq \frac{1}{2}$ , is given by

$$n \geq \frac{\ln 2}{\ln(1 + \delta)} \approx \frac{\ln 2}{\delta} = \frac{0.693}{\delta}.$$

Hence if  $\delta = 10^{-6}$ , we would need 693000 iterations to reach  $\frac{1}{2}$ . On the other hand, the bisection method would still converge to a relative accuracy of  $2^{-48}$  in 48 iterations.

We finish these short notes with some remarks on the error propagation. In actual computations, the exact value of  $g(x)$  may not be known. In general we can say that instead of  $g(x)$  we compute  $G(x) = g(x) + \delta(x)$ . What frequently happens is that we have a bound on  $\delta(x)$ , namely

$$\forall x \in [\alpha - \rho, \alpha + \rho] \quad |\delta(x)| < \delta.$$

In this case we can say that the actual iteration scheme is

$$x_{k+1} = g(x_k) + \delta_k \quad |\delta_k| < \delta.$$

What can we say about the convergence? We have the following result.

**Theorem.** - Let  $x_o$  be any point in  $[\alpha - \rho_o, \alpha + \rho_o]$  where  $0 < \rho_o < \rho - \frac{\delta}{1-L}$ . Then all the iterates lie in the interval  $[\alpha - \rho, \alpha + \rho]$  and moreover

$$|\alpha - x_k| \leq \frac{\delta}{1-L} + L^k(\rho_o - \frac{\delta}{1-L}) \quad \blacksquare.$$

This tells us that we should stop when  $L^k \rho_o \approx \frac{\delta}{1-L}$  provided we have reasonable estimates of  $L, \delta, \rho_o$ .

**Example.** - Let us evaluate the root of  $x^2 - 2$  in the interval  $[1, 3]$ . We get the following table of results, where the performances of the different methods are evident:

Newton's Method		$x_o = 1$	$x_1 = 1.5$	$x_2 = 1.41$	$x_3 = 1.4142157$
Secant Method	$x_{-1} = 3$	$x_o = 1$	$x_1 = 1.25$	$x_2 = 1.44$	$x_3 = 1.4126$
Chord Method		$x_o = 1$	$x_1 = 1.5$	$x_2 = 1.4375$	$x_3 = 1.4209$
Regula Falsi	$x_{-1} = 3$	$x_o = 1$	$x_1 = 1.25$	$x_2 = 1.353$	$x_3 = 1.392$