COURSE 10

5. Numerical methods for solving nonlinear equations in $\mathbb{R}$

Let $f : \Omega \to \mathbb{R}$, $\Omega \subset \mathbb{R}$. Consider the equation

$$f(x) = 0, \quad x \in \Omega. \tag{1}$$

Example. Kepler's Equation: consider a two-body problem like a satellite orbiting the earth or a planet revolving around the sun. Kepler discovered that the orbit is an ellipse and the central body F (earth, sun) is in a focus of the ellipse. The speed of the satellite P is not uniform: near the earth it moves faster than far away. It is used Kepler's law to predict where the satellite will be at a given time. If we want to know the position of the satellite for t = 9 minutes, then we have to solve the equation $f(E) = E - 0.8sinE - 2\pi/10 = 0$.

We attach a mapping $F : D \to D$, $D \subset \Omega^n$ to this equation.

Let $(x_0, ..., x_{n-1}) \in D$. Using F and the numbers $x_0, x_1, ..., x_{n-1}$ we construct iteratively the sequence

$$x_0, x_1, ..., x_{n-1}, x_n, ... \tag{2}$$

with

$$x_i = F\left(x_{i-n}, ..., x_{i-1}\right), \quad i = n, \ldots . \tag{3}$$

The problem consists in choosing F and $x_0, ..., x_{n-1} \in D$ such that the sequence (2) to be convergent to the solution of the equation (1).

Definition 1 *The procedure of approximation the solution of equation (1) by the elements of the sequence (2), computed as in (3), is called* F-**method.**

The numbers $x_0, x_1, ..., x_{n-1}$ *are called* **the starting points** *and the* k*-th element of the sequence (2) is called an approximation of* k*-th order of the solution.*

If the set of starting points has only one element then the F-method is **an one-step method;** if it has more than one element then the F-method is **a multistep method**.

Definition 2 *If the sequence (2) converges to the solution of the equation (1) then the F-method is convergent, otherwise it is divergent.*

Definition 3 *Let $\alpha \in \Omega$ be a solution of the equation (1) and let $x_0, x_1, ..., x_{n-1}, x_n, ...$ be the sequence generated by a given F-method. The number p having the property*

$$\lim_{x_i \to \alpha} \frac{\alpha - F(x_{i-n+1}, ..., x_i)}{(\alpha - x_i)^p} = C \neq 0, \quad C = constant,$$

is called the order *of the F-method.*

We construct some classes of F-methods based on the interpolation procedures.

Let $\alpha \in \Omega$ be a solution of the equation (1) and $V(\alpha)$ a neighborhood of α. Assume that f has inverse on $V(\alpha)$ and denote $g := f^{-1}$. Since

$$f(\alpha) = 0$$

it follows that

$$\alpha = g(0).$$

This way, the approximation of the solution α is reduced to the approximation of $g(0)$.

Definition 4 *The approximation of g by means of an interpolating method, and of α by the value of g at the point zero is called* **the inverse interpolation procedure.**

5.1. One-step methods

Let F be a one-step method, i.e., for a given x_i we have $x_{i+1} = F\left(x_i\right).$

Remark 5 *If* $p = 1$ *the convergence condition is* $\left|F'\left(x\right)\right| < 1$.

If $p > 1$ *there always exists a neighborhood of* α *where the* F*-method converges.*

All information on f are given at a single point, the starting value $\Rightarrow$ we are lead to Taylor interpolation.

Theorem 6 *Let* α *be a solution of equation (1),* $V(\alpha)$ *a neighborhood of* α*,* $x, x_i \in V(\alpha)$*,* f *fulfills the necessary continuity conditions. Then we have the following method, denoted by* F_m^T*, for approximating* α*:*

$$F_m^T(x_i) = x_i + \sum_{k=1}^{m-1} \frac{(-1)^k}{k!}[f(x_i)]^k g^{(k)}(f(x_i)), \tag{4}$$

where $g = f^{-1}$.

Proof. There exists $g = f^{-1} \in C^m[V(0)]$. Let $y_i = f(x_i)$ and consider Taylor interpolation formula

$$g(y) = (T_{m-1}g)(y) + (R_{m-1}g)(y),$$

with

$$(T_{m-1}g)(y) = \sum_{k=0}^{m-1} \tfrac{1}{k!}(y - y_i)^k g^{(k)}(y_i),$$

and $R_{m-1}g$ is the corresponding remainder.

Since $\alpha = g(0)$ and $g \approx T_{m-1}g$, it follows

$$\alpha \approx (T_{m-1}g)(0) = x_i + \sum_{k=1}^{m-1} \frac{(-1)^k}{k!} y_i^k g^{(k)}(y_i).$$

Hence,

$$x_{i+1} := F_m^T(x_i) = x_i + \sum_{k=1}^{m-1} \frac{(-1)^k}{k!}[f(x_i)]^k g^{(k)}(f(x_i))$$

is an approximation of α, and F_m^T is an approximation method for the solution α. $\blacksquare$

Concerning the order of the method F_m^T we state:

Theorem 7 *If* $g = f^{-1}$ *satisfies condition* $g^{(m)}(0) \neq 0$, *then* $\mathrm{ord}(F_m^T) = m$.

Remark 8 *We have an upper bound for the absolute error in approximating* α *by* x_{i+1} :

$$\left|\alpha - F_m^T(x_i)\right| \leq \frac{1}{m!}[f(x_i)]^m M_m g, \quad \textit{with } M_m g = \max_{y \in V(0)} \left|g^{(m)}(y)\right|.$$

Particular cases.

1) Case $m = 2$.

$$F_2^T\,(x_i) = x_i - \tfrac{f(x_i)}{f\,'(x_i)}.$$

This method is called **Newton's method** (**the tangent method**). Its order is 2.

2) Case $m = 3$.

$$F_3^T\,(x_i) = x_i - \tfrac{f(x_i)}{f\,'(x_i)} - \tfrac{1}{2}\left[\tfrac{f(x_i)}{f\,'(x_i)}\right]^2 \tfrac{f\,''(x_i)}{f\,'(x_i)},$$

with $\mathrm{ord}(F_3^T) = 3$. So, this method converges faster than F_2^T.

3) Case $m = 4$.

$$F_4^T(x_i) = x_i - \frac{f(x_i)}{f\,'(x_i)} - \frac{1}{2}\frac{f''(x_i)f^2(x_i)}{[f\,'(x_i)]^3} + \frac{\left(f'''(x_i)f'(x_i) - 3[f\,''(x_i)]^2\right)f^3(x_i)}{3![f\,'(x_i)]^5}.$$

Remark 9 *The higher the order of a method is, the faster the method converges. Still, this doesn't mean that a higher order method is more efficient (computation requirements). By the contrary, the most efficient are the methods of relatively low order, due to their low complexity (methods F_2^T and F_3^T).*

5.1.1. Newton's method

According to Remark 5, there always exists a neighborhood of α where the $F-$method is convergent. Choosing x_0 in such a neighborhood allows approximating α by terms of the sequence

$$x_{i+1} = F_2^T(x_i) = x_i - \frac{f(x_i)}{f'(x_i)}, \quad i = 0, 1, ...,$$

with a prescribed error ε.

If α is a solution of equation (1) and $x_{n+1} = F_2^T(x_n)$, for approximation error, Remark 8 gives

$$\left|\alpha - x_{n+1}\right| \leq \frac{1}{2}[f(x_n)]^2 M_2 g.$$

Lemma 10 *Let* $\alpha \in (a,b)$ *be a solution of equation (1) and let* $x_n = F_2^T(x_{n-1})$. *Then*

$$|\alpha - x_n| \leq \frac{1}{m_1}|f(x_n)|, \quad \textit{with } m_1 \leq m_1 f = \min_{a\leq x\leq b}\left|f'(x)\right|.$$

Proof. We use the mean formula

$$f(\alpha) - f(x_n) = f'(\xi)(\alpha - x_n),$$

with $\xi \in$ to the interval determined by α and $x_{n.}$ From $f(\alpha) = 0$ and $|f'(x)| \geq m_1$ for $x \in (a,b)$, it follows $|f(x_n)| \geq m_1|\alpha - x_n|$, that is

$$|\alpha - x_n| \leq \frac{1}{m_1}|f(x_n)|.$$

■

In practical applications the following evaluation is more useful:

Lemma 11 *If* $f \in C^2[a,b]$ *and* F_2^T *is convergent, then there exists* $n_0 \in \mathbb{N}$ *such that*

$$|x_n - \alpha| \leq |x_n - x_{n-1}|, \quad n > n_0.$$

Proof. We start with Taylor formula

$$f(x_n) = f(x_{n-1}) + (x_n - x_{n-1}) f'(x_{n-1}) + \tfrac{1}{2}(x_n - x_{n-1})^2 f''(\xi),$$

where ξ belongs to the interval determined by x_{n-1} and x_n.

Since $x_n = F_2^T(x_{n-1})$, it follows that

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})} \Longleftrightarrow f(x_{n-1}) + (x_n - x_{n-1}) f'(x_{n-1}) = 0,$$

thus we obtain

$$f(x_n) = \tfrac{1}{2}(x_n - x_{n-1})^2 f''(\xi).$$

Consequently,

$$|f(x_n)| \leq \tfrac{1}{2}(x_n - x_{n-1})^2 M_2 f,$$

and Lemma 10 yields $|\alpha - x_n| \leq \frac{1}{m_1} |f(x_n)|$ so

$$|\alpha - x_n| \leq \frac{1}{2m_1} (x_n - x_{n-1})^2 M_2 f.$$

Since F_2^T is convergent, there exists $n_0 \in \mathbb{N}$ such that

$$\frac{1}{2m_1} |x_n - x_{n-1}| M_2 f < 1, \quad n > n_0.$$

Hence,

$$|\alpha - x_n| \leq |x_n - x_{n-1}|, \quad n > n_0.$$

■

Remark 12 *The starting value is chosen randomly. If, after a fixed number of iterations the required precision is not achieved, i.e., condition* $|x_n - x_{n-1}| \leq \varepsilon$*, does not hold for a prescribed positive* ε*, the computation has to be started over with a new starting value.*

A modified form of Newton's method: - the same value during the computation of f':

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_0)}, \quad k = 0, 1, \ldots.$$

It is very useful because it doesn't request the computation of f' at x_j, $j = 1, 2, \ldots$ but the order is no longer equal to 2.

Another way for obtaining Newton's method.

We start with x_0 as an initial guess, sufficiently close to the α. Next approximation x_1 is the point at which the tangent line to f at $(x_0, f(x_0))$ crosses the Ox-axis. The value x_1 is much closer to the root α than x_0.

We write the equation of the tangent line at $(x_0, f(x_0))$:

$$y - f(x_0) = f'(x_0)(x - x_0).$$

If $x = x_1$ is the point where this line intersects the Ox-axis, then $y = 0$

$$-f(x_0) = f'(x_0)(x_1 - x_0),$$

and solving for x_1 gives

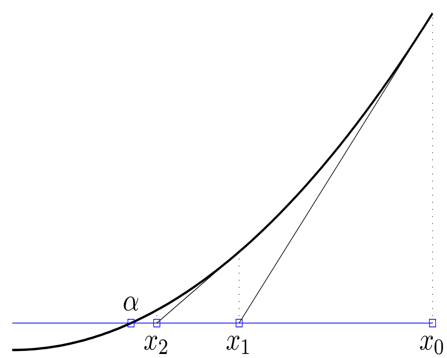
$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

By repeating the process using the tangent line at $(x_1, f(x_1))$, we obtain for x_2

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

For the general case we have

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \;\; n \geq 0. \tag{5}$$



The algorithm:

Let x_0 be the initial approximation.

for $n = 0, 1, ..., ITMAX$

$$x_{n+1} \leftarrow x_n - \frac{f(x_n)}{f'(x_n)}.$$

A stopping criterion is:

$$|f(x_n)| \leq \varepsilon \text{ or } \left|x_{n+1} - x_n\right| \leq \varepsilon \text{ or } \frac{\left|x_{n+1} - x_n\right|}{\left|x_{n+1}\right|} \leq \varepsilon,$$

where ε is a specified tolerance value.

Example 13 *Use Newton's method to compute a root of* $x^3 - x^2 - 1 = 0$, *to an accuracy of* 10^{-4}. *Use* $x_0 = 1$.

Sol. *The derivative of f is* $f'(x) = 3x^2 - 2x$. *Using* $x_0 = 1$ *gives* $f(1) = -1$ *and* $f'(1) = 1$ *and so the first Newton's iterate is*

$$x_1 = 1 - \frac{-1}{1} = 2 \text{ and } f(2) = 3,\ f'(2) = 8.$$

The next iterate is

$$x_2 = 2 - \frac{3}{8} = 1.625.$$

Continuing in this manner we obtain the sequence of approximations which converges to 1.465571.

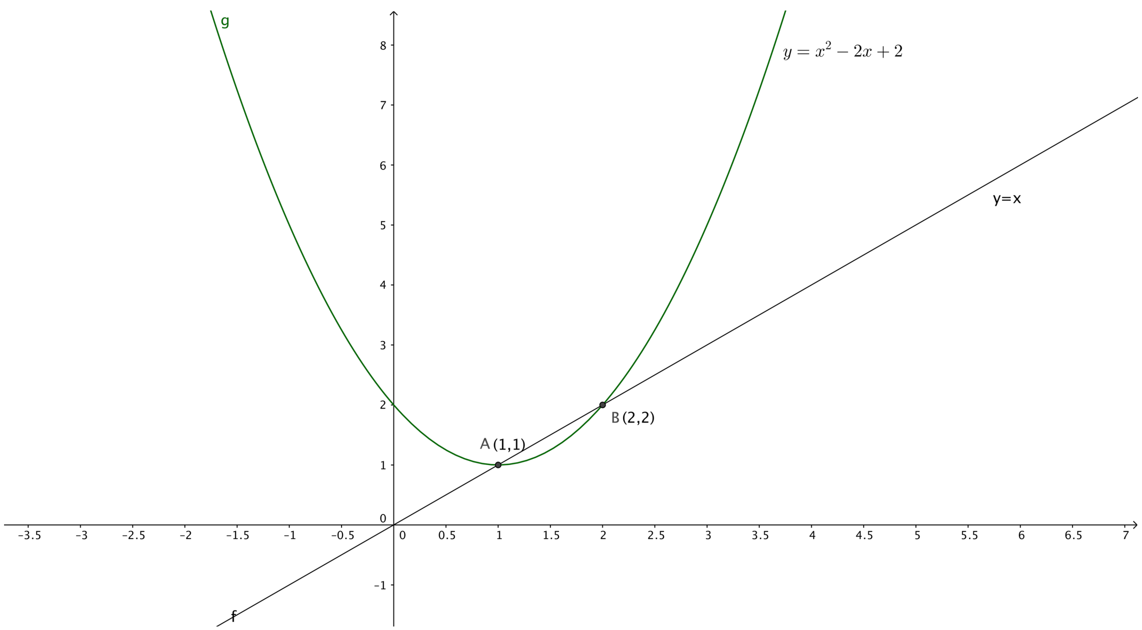
5.1.2. Fixed point iteration method (successive approximation method)

Definition 14 *The number α is called* **a fixed point** *of the function g if $g(\alpha) = \alpha$.*

Example 15 *Find the fixed points of the function $g(x) = x^2 - 2x + 2$.*

Sol. A fixed point α of g has the property $\alpha = g(\alpha) = \alpha^2 - 2\alpha + 2$, so $0 = \alpha^2 - 3\alpha + 2 = (\alpha - 1)(\alpha - 2)$. Whence, the fixed points of g are $\alpha_1 = 1$ and $\alpha_2 = 2$.

Geometrically, the fixed points are the intersection points of the graph of the function g and the first bisection line ($y = x$). (See the following figure.)



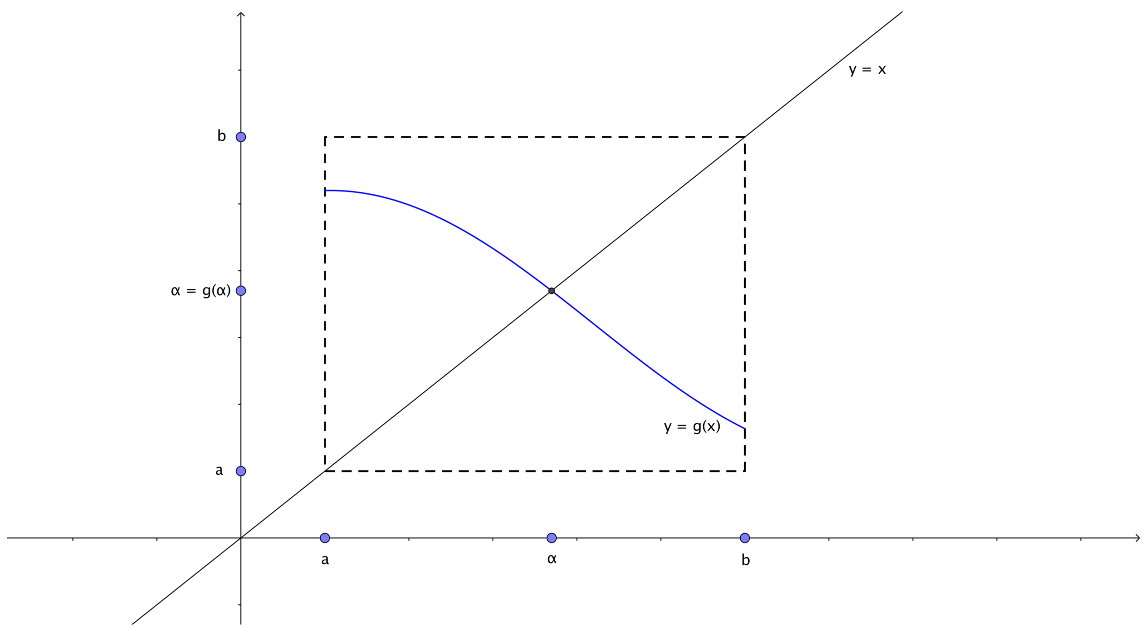
Sufficient condition for the existence and uniqueness of a fixed point:

Theorem 16 *1. If $g \in C[a,b]$ and $g(x) \in [a,b]$ for any $x \in [a,b]$, then g has at least one fixed point in $[a,b]$. In fewer words, if $g : [a,b] \to [a,b]$ and $g \in C[a,b]$ then $\exists \alpha \in [a,b]$ fixed point.*

2. Moreover, if there exists $g'(x)$ in (a,b) and

$$|g'(x)| < 1, \quad \forall x \in (a,b),$$

then the fixed point is unique in $[a,b]$.



Example 17 *Prove that $g(x) = (x^2 - 4)/5$ has a unique fixed point in $[-2, 2]$.*

Sol. The minimum and maximum of $g(x)$ for $x \in [-2, 2]$ are the limits of the interval, or at the points where $g'(x) = 0$. We have $g'(x) = 2x/5$, g is continuous and there exists $g'(x)$ in $[-2, 2]$. So, the minimum and maximum of $g(x)$ on $[-2, 2]$ are at $x = -2$, $x = 0$ or $x = 2$. We have $g(-2) = 0$, $g(2) = 0$, $g(0) = -4/5$, so $x = -2$ and $x = 2$ are points of absolute maximum and $x = 0$ is a point of absolute minimum in $[-2, 2]$. Moreover,

$$|g'(x)| = \left|\frac{2x}{5}\right| \le \left|\frac{4}{5}\right| < 1, \quad \forall x \in (-2, 2).$$

So, g satisfies the conditions of Theorem 16, so it follows that g has a unique fixed point in $[-2, 2]$.

Consider the equation

$$f(x) = 0, \tag{6}$$

where $f : [a, b] \to \mathbb{R}$. Assume that $\alpha \in [a, b]$ is a zero of $f(x)$.

In order to compute α, we transform (6) algebraically into *fixed point form*,

$$x = F(x), \tag{7}$$

where F is chosen so that $F(x) = x \Leftrightarrow f(x) = 0$.

A simple way to do this is, for example, $x = x + f(x) =: F(x)$.

Finding a zero of $f(x)$ in $[a, b]$ is then equivalent to finding a fixed point $x = F(x)$ in [a, b].

The fixed point form suggests *the fixed point iteration*

$$x_0 - \text{initial guess}, x_{k+1} = F(x_k),\ \ k = 0, 1, 2, \ldots.$$

The hope is that iteration will produce a convergent sequence $(x_n) \to \alpha$.

For example, consider

$$f(x) = xe^x - 1 = 0. \tag{8}$$

A first fixed point iteration is obtained rearranging and dividing (8) by e^x: $xe^x = 1 \Rightarrow x = e^{-x}$, so $x = F(x) = e^{-x}$ and

$$x_{k+1} = e^{-x_k}.$$

With the initial guess $x_0 = 0.5$ we obtain the iterates $x_1 = 0.6065306597$, $x_2 = 0.5452392119, ..., x_8 = 0.5664094527, x_9 = 0.5675596343, ...$, $x_{28} = 0.56714328, x_{29} = 0.56714329$

So x_k seems to converge to $\alpha = 0.5671432...$

A second fixed point form is obtained from $xe^x = 1$ by adding x on both sides: $xe^x + x = 1 + x \Rightarrow x(e^x + 1) = 1 + x \Rightarrow x = \frac{1+x}{e^x+1}$, we get

$$x = F(x) = \frac{1 + x}{e^x + 1}.$$

This time the convergence is much faster (we need only three iterations to obtain a 10-digit approximation of α) : $x_0 = 0.5$, $x_1 = 0.5663110032$, $x_2 = 0.5671431650$, $x_3 = 0.5671432904$.

Another possibility for a fixed point iteration is $x = x + 1 - xe^x$. But this iteration function does not generate a convergent sequence.

Finally we could also consider the fixed point form $x = x + xe^x - 1$. Also this iteration function does not generate a convergent sequence.

The question is: when does the iteration sequence converge?

Answer: when conditions of Theorem 16 are fulfilled.

For this example, we have two cases when $|F'(x)| < 1$ and the algorithm converges and two cases when $|F'(x)| > 1$ and the algorithm is not convergent.

A more general statement for the convergence is the theorem of Banach.

Definition 18 **A Banach space** *$\mathcal{B}$ is a complete normed vector space over some number field K such as $\mathbb{R}$ or $\mathbb{C}$.* (Complete *means that every Cauchy sequence converges in $\mathcal{B}$.)*

Definition 19 *Let $A \subset \mathcal{B}$ be a closed subset and $F : A \to A$. F is called* **Lipschitz continuous** *on A if there exists a constant $L \geq 0$ such that $||F(x) - F(y)|| \leq L\,||x - y||\,,\ \forall x, y \in A$. Furthermore, F is called* **a contraction** *if L can be chosen such that $L < 1$.*

Theorem 20 *(Banach Fixed Point Theorem) Let A be a closed subset of a Banach space $\mathcal{B}$, and let F be* **a contraction** *$F : A \to A$. Then:*

a) F has a unique fixed point α, which is the unique solution of the equation $x = F(x)$.

b) The sequence $x_{n+1} = F(x_n)$ *converges to* α *for every initial guess* $x_0 \in A$.

c) We have the estimate: $||\alpha - x_n|| \leq \frac{L^{n-l}}{1-L}||x_{l+1} - x_l||$, *for* $0 \leq l \leq n$ *(or* $||\alpha - x_n|| \leq \frac{L^n}{1-L}||x_1 - x_0||$)

For practical applications is also useful the following estimation.

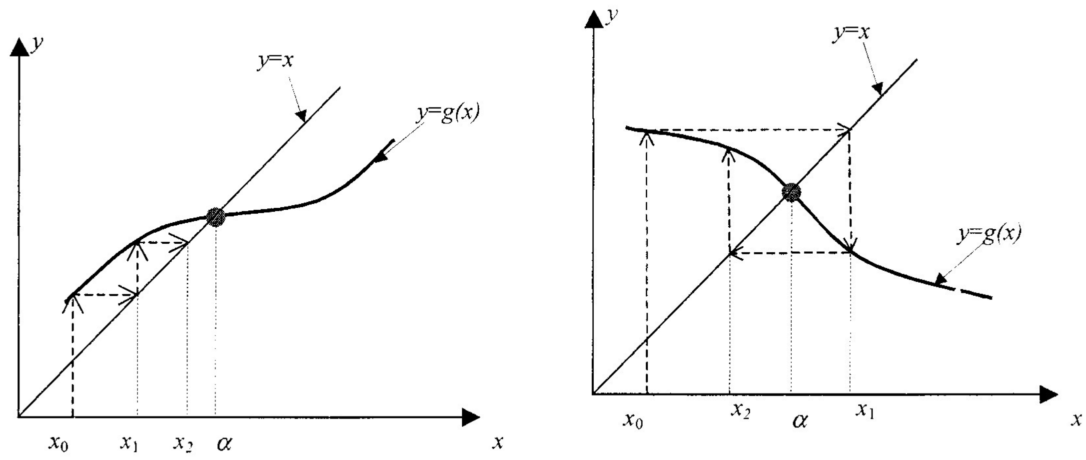
Lemma 21 *If* $\|F'(x)\| < L$, $x \in V(\alpha)$ *then*

$$||\alpha - x_n|| \leq \frac{L}{1-L}||x_n - x_{n-1}||.$$

Geometric interpretation of the method: we plot $y = F(x)$ and $y = x$. The intersection points of the two functions are the solutions of $x = F(x)$. The computation of the sequence $\{x_k\}$ with x_0 chosen initial value, $x_{k+1} = F(x_k), k = 0, 1, 2, ...$ can be interpreted geometrically via sequences of lines parallel to the coordinate axes:

x_0	start with x_0 on the x-axis
$F(x_0)$	go parallel to the y-axis to the graph of F
$x_1 = F(x_0)$	move parallel to the x-axis to the graph $y = x$
$F(x_1)$	go parallel to the y-axis to the graph of F
$etc.$	

Case of convergence $|F'(x)| < 1$.



Case of divergence $|F'(x)| > 1$.