# COURSE 10

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

*Example of nonlinear equation*. 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) 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 $f(E) = E - 0.8 \sin E - 2\pi/10 = 0$.

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

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

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 (5).

**Definition 1** *The procedure of approximation the solution of equation (5) 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 (initial) points** *and the* $k$*-th element of the sequence (2) is called an approximation of* $k$*-th index 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 (5) then the $F$-method is convergent, otherwise it is divergent.*

**Definition 3** *Let $\alpha \in \Omega$ be a solution of the equation (5) 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*

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

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

For one-step methods, the above condition reduces to

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

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

Let $\alpha \in \Omega$ be a solution of the equation (5) and $V(\alpha)$ a neighborhood of $\alpha$. 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 $\alpha$ is reduced to the approximation of $g(0)$.

**Definition 4** *The approximation of $g$ by means of an interpolating method, and of $\alpha$ 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 $x_{i+1} = F(x_i)$.

**Remark 5** *If $p = 1$, a sufficient convergence condition is $|F'(x)| < 1$.*

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

**Theorem 6** *Let $\alpha$ be a solution of equation (5), $V(\alpha)$ a neighborhood of $\alpha$, $x, x_i \in V(\alpha)$, $f$ fulfills the necessary continuity conditions. Then we have the following method, denoted by $F_m^T$, for approximating $\alpha$:*

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

*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

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

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 $\alpha$, and $F_m^T$ is an approximation method for the solution $\alpha$. ■

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

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

**Remark 8** *We have an upper bound for the absolute error in approximating* $\alpha$ *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 $m = 2$.

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

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

2) Case $m = 3$.

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

with $\text{ord}(F_3^T) = 3$. 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 (due to 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* $F_3^T$*).*

Newton's method (Newton-Raphson method) named after Isaac Newton (1642–1726) and Joseph Raphson (1648–1715), is a root-finding algorithm which produces successively better approximations to the roots of a real-valued function.

This method is so efficient in computing $\sqrt{a}$, that it is a choice even today in modern codes.

**Some comments on the history of this method**

The traces of this methods can be found in ancient times (Babylon and Egypt, 1800 B.C.), as it appears in the computation of the square root of a number.

Different methods, either algebraically equivalent to Newton's method or of Newton type were known in antiquity to India and then to Arabic culture.

In solving nonlinear problems, Newton ($\approx$1669) and subsequently Raphson (1690) have dealt only with polynomial equations ($x^3-2x-5=0$ is "the classical equation where the Newton method is applied"). Newton has also considered such iterations in solving Kepler's equation $x - e\sin x = M$.

Newton has considered the process of successively computing the corrections, which were added finally altogether to form the final approximation. He didn't compute the successive approximations, but

computes a sequence of polynomials, and only at the end arrives at an approximation for the root: let $x_0$ be a given first estimate of the solution $\alpha$ of $f(x) = 0$. Write $g_0(x) = f(x)$, and suppose $g_0(x) = \sum_{i=0}^{n} a_i x^i$. Writing $e_0 = \alpha - x_0$ we obtain by binomial expansion about the given $x_0$ a new polynomial equation in the variable $e_0$:

$$0 = g_0(\alpha) = g_0(x_0 + e_0) = \sum_{i=0}^{n} a_i (x_0 + e_0)^i$$
$$= \sum_{i=0}^{n} a_i \left[ \sum_{j=0}^{i} \binom{i}{j} x_0^j {e_0}^{i-j} \right] = g_1(e_0).$$

Neglecting terms involving higher powers of $e_0$ (linearizing the explicitly computed polynomial $g_1$) produces

$$0 = g_1(e_0) \approx \sum_{i=0}^{n} a_i (x_0^i + i x_0^{i-1} e_0) = \sum_{i=0}^{n} a_i x_0^i + e_0 \sum_{i=0}^{n} a_i i x_0^{i-1}$$

from which we deduce that

$$e_0 \approx c_0 = \left( -\sum_{i=0}^{n} a_i x_0^i \right) \Big/ \left( \sum_{i=0}^{n} a_i i x_0^{i-1} \right)$$

and set $x_1 = x_0 + c_0$. Formally this correction can be written $c_0 = -g_0(x_0)/g_0'(x_0) = -f(x_0)/f'(x_0)$.

Now repeat the process, but instead of expanding the original polynomial $g_0$ about $x_1$ expand the polynomial $g_1$ about the point $c_0$, that is considered to be a first estimate of the solution $e_0$ of the new equation $g_1(x) = 0$. Thus similarly obtain $0 = g_1(e_0) = g_1(c_0 + e_1) = g_2(e_1)$, where the polynomial $g_2$ is explicitly computed. Liniarizing again produces $e_1 \approx c_1 = -g_1(c_0)/g_0'(c_0)$, corresponding $x_2 = x_1 + c_1$.

Raphson has considered the approximations updated at each step (the usual iterations), a process equivalent to what we use nowadays. However, the derivatives of $f$ (which could be calculated with the "fluxions" of that time) do not appear in their formulae. For more than a century, the belief was that these two variants (of Newton and Raphson) represent two different methods.

Simpson (1740) was the first to apply the method to transcendent equations, using the "fluxions" (the fluxions $\dot{x}$ are "essentially" equivalent to $dx/dt$.) He even extended it to the solving of nonlinear systems

of two equations, subsequently generalized to the usual form from today.

The formulation of the method using $f'(x)$ was published by Lagrange in 1798.

Due to the Fourier's influential book *Analyse des Équations Determinées* (1837), where Raphson and Simpson were not mentioned, the name of the method remained "Newton". Some authors use "the Newton-Raphson method".

When there exists a neighborhood of $\alpha$ where the $F$-method is convergent, choosing $x_0$ in such a neighborhood allows approximating $\alpha$ 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 $\varepsilon$.

If $\alpha$ is a solution of equation (5) and $x_{n+1} = F_2^T(x_n)$, for approximation error, Remark 8 gives

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

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

$$f(x) = 0, \quad x \in \Omega. \tag{5}$$

For $m = 2$.

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

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

When there exists a neighborhood of $\alpha$ where the $F$-method is convergent, choosing $x_0$ in such a neighborhood allows approximating $\alpha$ by terms of the sequence

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

with a prescribed error $\varepsilon$.

**Lemma 10** *Let $\alpha \in (a,b)$ be a solution of equation (5) and let $x_n = F_2^T(x_{n-1})$. Then*

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

**Proof.** We use the mean formula

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

with $\xi \in$ to the interval determined by $\alpha$ 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 $\xi$ 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* $\varepsilon$, *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.