Numerical methods

Session 3: Rootfinding of nonlinear equations

Pedro González Rodríguez

Universidad Carlos III de Madrid

February 12, 2020

Problem: Given $f : I = (a, b) \subseteq \mathbb{R} \to \mathbb{R}$, find $\alpha \in \mathbb{C}$ such that $f(\alpha) = 0$.

Definition: A sequence $\left\{x^{(k)}\right\}$ generated by a numerical method is said to converge to α with order $p \geq 1$ if

$$\exists C > 0 : \frac{|x^{(k+1)} - \alpha|}{|x^{(k)} - \alpha|^p} \leq C, \forall k \geq k_0,$$

where $k_0 \geq 0$ is a suitable integer. In such a case, the method is said to be of order p. Notice that if p is equal to 1, in order for $x^{(k)}$ to converge to α it is necessary that $C < 1$. In such an event, the constant C is called the convergence factor of the method.

Rootfinding of nonlinear equations

The convergence of iterative methods for rootfinding of nonlinear equations depends in general on the choice of the initial datum $x^{(0)}$. This allows for establishing only local convergence results, that is, holding for any $x^{(0)}$ which belongs to a suitable neighborhood of the root α. Methods for which convergence to α holds for any choice of $x^{(0)}$ in the interval I, are said to be globally convergent to α.

Conditioning of a nonlinear equation

Consider $f(x) = \varphi(x) - d = 0$ with $f \in C^\infty$. Then:

- The problem is well posed only if the function φ is invertible. In such a case one gets $\alpha = \varphi^{-1}(d) = G(d)$.
- $$K(d) \approx \frac{|d|}{|\alpha||f'(\alpha)|}, \qquad K_{abs}(d) \approx \frac{1}{|f'(\alpha)|} \tag{1}$$

The problem is thus ill-conditioned when $f'(\alpha)$ is "small" and well-conditioned if $f'(\alpha)$ is "large".

Conditioning of a nonlinear equation

The analysis can be generalized to roots α with multiplicity $m > 1$. Expanding φ in a Taylor series around α up to the m-th order term, we get

$$d + \delta d = \varphi(\alpha + \delta\alpha) = \varphi(\alpha) + \sum_{k=1}^{m} \frac{\varphi^{(k)}(\alpha)}{k!}(\delta\alpha)^k + o((\delta\alpha)^m).$$

Since $\varphi^{(k)}(\alpha) = 0$ for $k = 1, ..., m-1$, we obtain $\delta d = f^{(m)}(\alpha)(\delta\alpha)^m/m!$ so that an approximation to the absolute condition number is

$$K_{abs}(d) \approx \left| \frac{m!\delta d}{f^{(m)}(\alpha)} \right|^{1/m} \frac{1}{|\delta d|}. \tag{2}$$

Conditioning of a nonlinear equation

Notice that the expression for the condition number of a simple root is a special case of this one using $m = 1$. From this it also follows that, even if δd is sufficiently small to make $\left|(m!\delta d)/(f^{(m)}(\alpha))\right| < 1$, $K_{abs}(d)$ could nevertheless be a large number. We therefore conclude that the problem of rootfinding of a nonlinear equation is well-conditioned if α is a simple root and $|f'(\alpha)|$ is definitely different from zero, ill-conditioned otherwise.

Conditioning of a nonlinear equation

Consider the following problem: Assume $d = 0$, let α be a simple root of f and $\hat{\alpha} \neq \alpha$, let $f(\hat{\alpha}) = \hat{r} \neq 0$. We seek a bound for the difference $\hat{\alpha} - \alpha$ as a function of the residual $\hat{r}$. Applying (1) yields

$$K_{abs}(0) \approx \frac{1}{|f'(\alpha)|}.$$

Therefore, letting $\delta x = \hat{\alpha} - \alpha$ and $\delta d = \hat{r}$ in the definition of K_{abs},($K_{abs} = \sup_{\delta d \in D} \frac{\|\delta x\|}{\|\delta d\|}$), we get

$$\frac{|\hat{\alpha} - \alpha|}{|\alpha|} \lesssim \frac{|\hat{r}|}{|f'(\alpha)||\alpha|}.$$

Conditioning of a nonlinear equation

If α has multiplicity $m > 1$, using (2) instead of (1) and proceeding as above, we get

$$\frac{|\hat{\alpha} - \alpha|}{|\alpha|} \lesssim \left(\frac{m!}{|f^{(m)}(\alpha)||\alpha|^m} \right)^{1/m} |\hat{r}|^{1/m}. \tag{3}$$

These estimates will be useful in the analysis of stopping criteria for iterative methods.

A geometric approach to root finding: Bisection method

The bisection method is based on the following property.

Property 6.1 (theorem of zeros for continuous functions): Given a continuous function $f : [a, b] \to \mathbb{R}$, such that $f(a)f(b) < 0$, then $\exists \alpha \in (a, b)$ such that $f(\alpha) = 0$.

Starting from $I_0 = [a, b]$, the bisection method generates a sequence of subintervals $I_k = [a(k), b(k)], k \geq 0$, with $I_k \subset I_{k-1}, k \geq 1$, fulfilling the property that $f(a(k))f(b(k)) < 0$. Precisely, we set $a(0) = a$, $b(0) = b$ and $x^{(0)} = (a(0) + b(0))/2$.

A geometric approach to root finding: Bisection method

The bisection iteration terminates at the m-th step for which $|x(m) - \alpha| \leq |I_m| \leq \epsilon$, where ϵ is a fixed tolerance and $|I_m|$ is the length of I_m. As for the speed of convergence of the bisection method, notice that $|I_0| = b - a$, while

$$|I_k| = |I_0|/2^k = (b-a)/2^k, k \geq 0.$$

A geometric approach to root finding: Bisection method

Denoting by $e^{(k)} = x^{(k)} - \alpha$ the absolute error at step k, it follows that $|e^{(k)}| \leq (b-a)/2^k, k \geq 0$, which implies $\lim_{k \to \infty} |e^{(k)}| = 0$. The bisection method is therefore globally convergent.

A geometric approach to root finding: Bisection method

Moreover, to get $|x^{(m)} - \alpha| \leq \epsilon$ we must take

$$m \geq log2(b-a) - log2(\epsilon) = \frac{log((b-a)/\epsilon)}{log(2)} \approx \frac{log((b-a)/\epsilon)}{0.6931}.$$

In particular, to gain a significant figure in the accuracy of the approximation of the root (that is $|x^{(k)} - \alpha| = |x^{(j)} - \alpha|/10$), one needs $k - j = log2(10) \approx 3.32$ bisections.

This singles out the bisection method as an algorithm of certain, but slow, convergence. We must also point out that the bisection method does not generally guarantee a monotone reduction of the absolute error between two successive iterations, that is, we cannot ensure a priori that

$$|e^{(k+1)}| \leq M_k |e^{(k)}| \text{ for any } k \geq 0$$

with $M_k < 1$. Failure to satisfy this does not allow for qualifying the bisection method as a method of order 1, according to the definition given in the previous section.

The Methods of Chord, Secant and Regula Falsi and Newton's Method

In order to devise algorithms with better convergence properties than the bisection method, it is necessary to include information from the values attained by f and, possibly, also by its derivative f' (if f is differentiable) or by a suitable approximation. For this purpose, let us expand f in a Taylor series around α and truncate the expansion at the first order. The following linearized version of problem $f : I = (a, b) \subseteq \mathbb{R} \to \mathbb{R}$, find $\alpha \in C$ such that $f(\alpha) = 0$ is obtained

$$f(\alpha) = 0 = f(x) + (\alpha - x)f'(\psi)$$

for a suitable ψ between α and x.

This equation prompts the following iterative method: for any $k \geq 0$, given $x^{(k)}$, determine $x^{(k+1)}$ by solving equation

$$f(x^{(k)}) + (x^{(k+1)} - x^{(k)})q_k = 0,$$

where q_k is a suitable approximation of $f'(x^{(k)})$. The method described here amounts to finding the intersection between the x-axis and the straight line of slope q_k passing through the point $(x^{(k)}, f(x^{(k)}))$, and thus can be more conveniently set up in the form

$$x^{(k+1)} = x^{(k)} - q_k^{-1} f(x^{(k)}), \forall k \geq 0.$$

The Methods of Chord, Secant and Regula Falsi and Newton's Method

We consider below four particular choices of q_k:
The chord method. We let

$$q_k = q = \frac{f(b) - f(a)}{b - a}, \forall k \geq 0$$

from which, given an initial value $x^{(0)}$, the following recursive relation is obtained

$$x^{(k+1)} = x^{(k)} - \frac{b - a}{f(b) - f(a)} f(x^{(k)}), k \geq 0.$$

The order of convergence of this method is $p = 1$.

The Secant method. We let

$$q_k = \frac{f(x^{(k)}) - f(x^{(k-1)})}{x^{(k)} - x^{(k-1)}}, \forall k \geq 0.$$

(6.13)

from which, giving two initial values $x^{(-1)}$ and $x^{(0)}$, we obtain the following relation

$$x^{(k+1)} = x^{(k)} - \frac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})} f(x^{(k)}), k \geq 0.$$

(6.14)

Requires an extra initial point, and the incremental ratio at each step.The order of convergence is $p = (1 + \sqrt{5})/2 \approx 1.63$.

The Regula Falsi (or false position) method. This is a variant of the secant method in which, instead of selecting the secant line through the values $(x^{(k)}, f(x^{(k)}))$ and $(x^{(k-1)}, f(x^{(k-1)}))$, we take the one through $(x^{(k)}, f(x^{(k)}))$ and $(x(k'), f(x(k')))$, k' being the maximum index less than k such that $f(x(k')) \cdot f(x^{(k)}) < 0$. Precisely, once two values $x^{(-1)}$ and $x^{(0)}$ have been found such that $f(x^{(-1)}) \cdot f(x^{(0)}) < 0$, we let

$$x^{(k+1)} = x^{(k)} - \frac{x^{(k)} - x^{(k')}}{f(x^{(k)}) - f(x^{(k')})} f(x^{(k)}), k \geq 0.$$

The Methods of Chord, Secant and Regula Falsi and Newton's Method

Notice that the sequence of indices k' is nondecreasing; therefore, in order to find at step k the new value of k', it is not necessary to sweep all the sequence back, but it suffices to stop at the value of k' that has been determined at the previous step.
The Regula Falsi method, though of the same complexity as the secant method, has linear convergence order. However, unlike the secant method, the iterates generated by the method are all contained within the starting interval $[x^{(-1)}, x^{(0)}]$.
In this respect, the Regula Falsi method, as well as the bisection method, can be regarded as a globally convergent method.

The Newton's method. Assuming that $f \in C^1(I)$ and that $f'(\alpha) \neq 0$ (i.e., α is a simple root of f), if we let $q_k = f'(x^{(k)}), \forall k \geq 0$ and assign the initial value $x^{(0)}$, we obtain the so called Newton's method

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})}, k \geq 0.$$

At the k-th iteration, Newton's method requires the two functional evaluations $f(x^{(k)})$ and $f'(x^{(k)})$. The increasing computational cost with respect to the methods previously considered is more than compensated for by a higher order of convergence, Newton's method being of order 2.

Newton's Method for simultaneous nonlinear equations.

Let

$$\vec{F}(\vec{x}) = \begin{bmatrix} f_1(\vec{x}) \\ f_2(\vec{x}) \end{bmatrix} \approx$$

$$\begin{bmatrix} f_1(\vec{x}^{(1)}) + \frac{df_1}{dx_1}(\vec{x}^{(1)})(x_1 - x_1^{(1)}) + \frac{df_1}{dx_2}(\vec{x}^{(1)})(x_2 - x_2^{(1)}) \\ f_2(\vec{x}^{(1)}) + \frac{df_2}{dx_1}(\vec{x}^{(1)})(x_1 - x_1^{(1)}) + \frac{df_2}{dx_2}(\vec{x}^{(1)})(x_2 - x_2^{(1)}) \end{bmatrix}$$

Newton's Method for simultaneous nonlinear equations.

Assuming that $\vec{F}(\vec{x}) = 0$ and doing some calculations:

$$\begin{bmatrix} -f_1(\vec{x}^{(1)}) \\ -f_2(\vec{x}^{(1)}) \end{bmatrix} =$$

$$\begin{bmatrix} \frac{df_1}{dx_1}(\vec{x}^{(1)}) & \frac{df_1}{dx_2}(\vec{x}^{(1)}) \\ \frac{df_2}{dx_1}(\vec{x}^{(1)}) & \frac{df_2}{dx_2}(\vec{x}^{(1)}) \end{bmatrix} \begin{bmatrix} (x_1 - x_1^{(1)}) \\ (x_2 - x_2^{(1)}) \end{bmatrix}$$

Newton's Method for simultaneous nonlinear equations.

And then:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \end{bmatrix} - \begin{bmatrix} \frac{df_1}{dx_1}(\vec{x}^{(1)}) & \frac{df_1}{dx_2}(\vec{x}^{(1)}) \\ \frac{df_2}{dx_1}(\vec{x}^{(1)}) & \frac{df_2}{dx_2}(\vec{x}^{(1)}) \end{bmatrix}^{-1} \begin{bmatrix} f_1(\vec{x}^{(1)}) \\ f_2(\vec{x}^{(1)}) \end{bmatrix}$$

or

$$\vec{x}^{(i+1)} = \vec{x}^{(i)} - J(\vec{x}^{(i)})^{-1} \vec{F}(\vec{x}^{(i)})$$

Stopping Criteria.

Suppose that $\left\{x^{(k)}\right\}$ is a sequence converging to a zero α of the function f.
Below, ϵ is a fixed tolerance, $e^{(k)} = \alpha - x^{(k)}$ denotes the absolute error, and we assume that f is continuously differentiable in a suitable neighborhood of the root.

Stopping Criteria.

There are two possible stopping criteria:

- Stopping test based on the residual: The iterative process terminates at the first step k such that $|f(x^{(k)})| < \epsilon$.
- Stopping test based on the increment: the iterative process terminates as soon as $|x^{(k+1)} - x^{(k)}| < \epsilon$.

First test analysis: Situations can arise where the test turns out to be either too restrictive or excessively optimistic. Remember (3):

$$\frac{|e^{(k)}|}{|\alpha|} \lesssim \left(\frac{m!}{|f^{(m)}(\alpha)||\alpha|^m} \right)^{1/m} |f(x^{(k)})|^{1/m}.$$

In particular, in the case of simple roots, the error is bound to the residual by the factor $1/|f'(\alpha)|$ so that the following conclusions can be drawn:

Stopping Criteria.

1. If $|f'(\alpha)| \approx 1$, then $|e^{(k)}| \approx \epsilon$ therefore, the test provides a satisfactory indication of the error;
2. If $|f'(\alpha)| << 1$, the test is not reliable since $|e^{(k)}|$ could be quite large.
3. If, finally, $|f'(\alpha)| >> 1$, we get $|e^{(k)}| << \epsilon$ and the test is too restrictive with respect to ϵ.

Second test analysis. Let $\{x^{(k)}\}$ be generated by the fixed-point method $x^{(k+1)} = \phi(x^{(k)})$. Using the mean value theorem, we get

$$e^{(k+1)} = \phi(\alpha) - \phi(x^{(k)}) = \phi'(\xi^{(k)})e^{(k)},$$

where $\xi^{(k)}$ lies between $x^{(k)}$ and α. Then,

$$x^{(k+1)} - x^{(k)} = e^{(k)} - e^{(k+1)} = 1 - \phi'(\xi^{(k)})e^{(k)},$$

so that, assuming that we can replace $\phi'(\xi^{(k)})$ with $\phi'(\alpha)$, it follows that

$$e^{(k)} \approx \frac{1}{1 - \phi'(\alpha)}(x^{(k+1)} - x^{(k)}).$$

Stopping Criteria.

We can conclude that the test:

1. is unsatisfactory if $\phi'(\alpha)$ is close to 1.
2. provides an optimal balancing between increment and error in the case of methods of order 2 for which $\phi'(\alpha) = 0$ as is the case for Newton's method.
3. is still satisfactory if $-1 < \phi'(\alpha) < 0$.