Numerical Analysis: Solving Nonlinear Equations

Mirko Navara
http://cmp.felk.cvut.cz/˜navara/
Center for Machine Perception, Department of Cybernetics, FEE, CTU
Karlovo náměstí, building G, office 104a

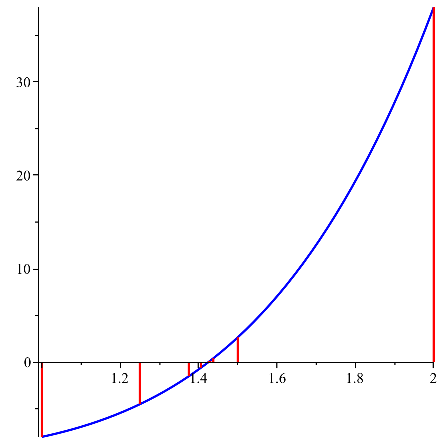
http://math.feld.cvut.cz/nemecek/nummet.html

November 10, 2014

Task:
Find a real solution to the equation $f(x) = 0$, where f is a continuous real function at interval $[a_0, b_0]$.
More specification is needed:
Task:
Find a real solution to the equation $f(x) = 0$, where f is a continuous real function at interval $[a_0, b_0]$.
We assume that $f(a_0) \cdot f(b_0) < 0$ (i.e., $f(a_0), f(b_0)$ have the opposite signs) and f has exactly one root, $\overline{x}$, at interval $[a_0, b_0]$. The required precision of the solution is $\varepsilon > 0$, i.e., we have to find a value in the interval $[\overline{x} - \varepsilon, \overline{x} + \varepsilon]$.
The first step is the **separation of roots**, which is not algorithmizable.

$$x_i = \frac{a_i + b_i}{2}$$

- if $f(x_i) \cdot f(a_i) < 0$, then $a_{i+1} = a_i$, $b_{i+1} = x_i$,
- if $f(x_i) \cdot f(b_i) < 0$, then $a_{i+1} = x_i$, $b_{i+1} = b_i$,
- if $f(x_i) = 0$, then $\overline{x} = x_i$.



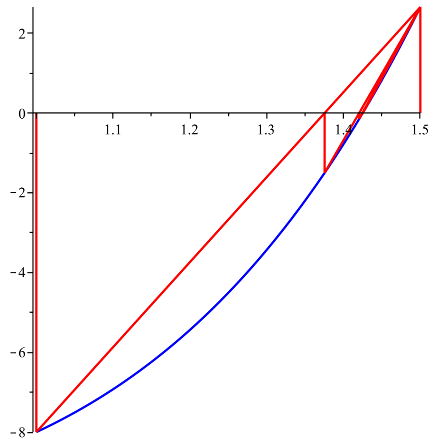
Ending condition: $\dfrac{b_i - a_i}{2} \le \varepsilon$

It always converges with a constant speed, approx. 3 decimal places in 10 iterations.

We divide the interval $[a_i, b_i]$ in proportion $\dfrac{|f(a_i)|}{|f(b_i)|}$:

$$\frac{x_i - a_i}{x_i - b_i} = \frac{f(a_i)}{f(b_i)}$$

$$x_i = \frac{a_i f(b_i) - b_i f(a_i)}{f(b_i) - f(a_i)}$$



Typically, one endpoint remains unchanged (e.g., if f'' does not change its sign).

$$b_i - a_i \not\to 0$$

$$\lim_{i\to\infty}(b_i - a_i) \in \{|\overline{x} - a_j|, |\overline{x} - b_j| : j \in \mathbb{N}_0\}$$

Ending condition: $|f(x_i)| \le \delta$
Evaluate the Taylor series of function f with center x_i at $\overline{x}$:

$$\underbrace{f(\overline{x})}_{0} = f(x_i) + (\overline{x} - x_i)\, f'(\theta_i)$$

for some $\theta_i \in I(x_i, \overline{x})$

$$\overline{x} - x_i = \frac{-f(x_i)}{f'(\theta_i)}$$

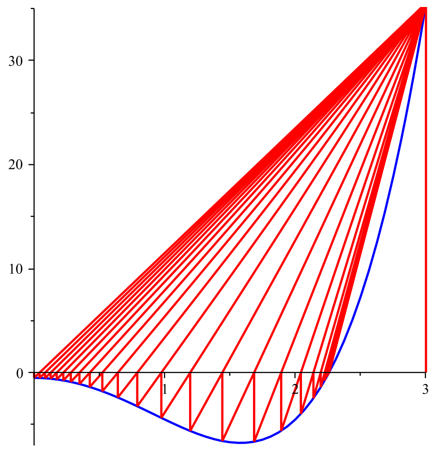
If $\exists m_1 > 0\ \forall x \in I(x_i, \overline{x}) : m_1 \le |f'(x)|$,
for absolute values we get

$$|\overline{x} - x_i| \le \frac{|f(x_i)|}{m_1}$$

Theorem: Let f have a continuous derivative at interval $I(x_i, \overline{x})$ and
$\exists m_1 > 0\ \forall x \in I(x_i, \overline{x}) : m_1 \le |f'(x)|$. Then

$$|\overline{x} - x_i| \le \frac{|f(x_i)|}{m_1} \le \frac{\delta}{m_1}$$

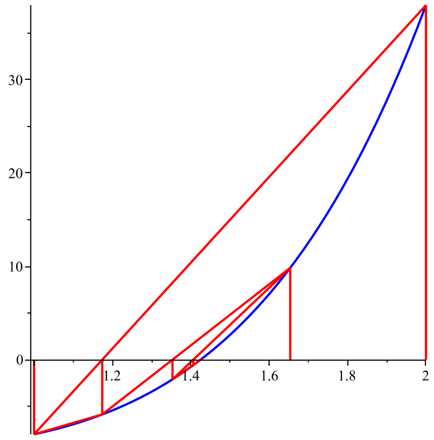
The theorem cannot be used if the derivative does not exist or the root is multiple. (The method can still be applicable.)
Regula falsi converges faster if the given function is approximately linear (in a neighborhood of the root).



Modification of regula falsi: We always continue with the last two points, independently of the signs:

$$x_0 = b_0, \qquad x_1 = a_0$$

$$x_i = \frac{x_{i-2} f(x_{i-1}) - x_{i-1} f(x_{i-2})}{f(x_{i-1}) - f(x_{i-2})}$$



Ending condition: $\quad |f(x_i)| \le \delta \quad$ or $\quad |x_i - x_{i-1}| \le \eta$
The convergence is usually faster, but it is not guaranteed.
Methods

- **single-point**
- **double-point**
- **multiple-point**

Tangent to the graph of function f at $\big(x_{i-1}, f(x_{i-1})\big)$:

$$t_{i-1}(x) = f(x_{i-1}) + (x - x_{i-1}) \cdot f'(x_{i-1})$$

x_i is its zero point (root):

$$x_i = x_{i-1} - \frac{f(x_{i-1})}{f'(x_{i-1})}$$

It assumes the existence of a **known** first derivative; it is necessary to fix overlow and division by zero.
Ending condition: $\quad |f(x_i)| \le \delta \quad$ or $\quad |x_i - x_{i-1}| \le \eta$
The convergence is usually faster, but it is not guaranteed.
Evaluate the Taylor series of function f with center x_{i-1} at x_i:

$$f(x_i) = \underbrace{f(x_{i-1}) + (x_i - x_{i-1})\, f'(x_{i-1})}_{0} + \frac{1}{2}(x_i - x_{i-1})^2\, f''(\xi_i)$$

where $\xi_i \in I(x_i, x_{i-1})$. Substitute in the universal estimate of the error:

$$\overline{x} - x_i = \frac{-f(x_i)}{f'(\theta_i)} = \frac{-f''(\xi_i)}{2f'(\theta_i)}\,(x_i - x_{i-1})^2$$

If there are estimates

$$\begin{aligned} \exists M_2\ \forall x \in I(x_i, \overline{x}) : &\qquad |f''(x)| \le M_2 \\ \exists m_1 > 0\ \forall x \in I(x_i, x_{i-1}) : &\qquad |f'(x)| \ge m_1 \end{aligned}$$

then we obtain (for absolute values)

$$|\overline{x} - x_i| \le \frac{M_2}{2m_1}\,(x_i - x_{i-1})^2$$

Theorem: Let $\overline{x}$ be a **simple** root of function f, which has a continuous second derivative at interval $I(x_i, x_{i-1}, \overline{x})$ (where x_i is the result of one step of Newton's method applied to the estimate x_{i-1}). Let there exist real numbers M_2, $m_1 > 0$ such that

$$\begin{aligned} \forall x \in I(x_i, \overline{x}): & \quad |f'(x)| \geq m_1 \\ \forall x \in I(x_i, x_{i-1}): & \quad |f''(x)| \leq M_2 \end{aligned}$$

Then we have an estimate of the error

$$|\overline{x} - x_i| \leq \frac{M_2}{2m_1} (x_i - x_{i-1})^2$$

Corollary: When the ending condition $|x_i - x_{i-1}| \leq \eta$ is satisfied, we get an estimate of the error

$$|\overline{x} - x_i| \leq \frac{M_2}{2m_1} \eta^2$$

Simple rule: If Newton's method converges and the approximation is near the root, the number of valid digits behind the decimal point doubles in each iteration.
Correction: If the error is much smaller than 1 and absolute values of the first two derivatives of f are approximately the same, then we may neglect the factor $\frac{M_2}{2m_1}$ and the latter rule is valid because

$$|\overline{x} - x_i| \approx (x_i - x_{i-1})^2 \approx (\overline{x} - x_{i-1})^2 .$$

However, if the proportion $\frac{M_2}{2m_1}$ is much different from 1, the rule cannot be used.
Not guaranteed, in particular for a bad initial estimate.
Assumption: Function f has a continuous second derivative in a neighborhood of a **simple** root $\overline{x}$.
Then $f'(\overline{x}) \neq 0$ and f' is continuous in a neighborhood of $\overline{x}$
$\Rightarrow$ we can find a neighborhood I of point $\overline{x}$ such that

$$\begin{aligned} \exists m_1 > 0 \ \forall x \in I: & \quad |f'(x)| \geq m_1 \\ \exists M_2 \ \forall x \in I: & \quad |f''(x)| \leq M_2 \end{aligned}$$

Let $x_{i-1} \in I \setminus \{\overline{x}\}$.
We evaluate the Taylor series of function f with center x_{i-1} at $\overline{x}$:

$$\underbrace{f(\overline{x})}_{0} = f(x_{i-1}) + (\overline{x} - x_{i-1}) f'(x_{i-1}) + \tfrac{1}{2} (\overline{x} - x_{i-1})^2 f''(\xi_i) ,$$

where $\xi_i \in I(\overline{x}, x_{i-1})$. Subtract

$$\begin{aligned} 0 &= f(x_{i-1}) + (x_i - x_{i-1}) f'(x_{i-1}) \\ 0 &= (\overline{x} - x_i) f'(x_{i-1}) + \tfrac{1}{2} (\overline{x} - x_{i-1})^2 f''(\xi_i) \\ \frac{\overline{x} - x_i}{(\overline{x} - x_{i-1})^2} &= -\frac{f''(\xi_i)}{2f'(x_{i-1})} \\ \frac{|\overline{x} - x_i|}{(\overline{x} - x_{i-1})^2} &\leq \frac{M_2}{2m_1} \\ \frac{|\overline{x} - x_i|}{|\overline{x} - x_{i-1}|} &\leq \frac{M_2}{2m_1} |\overline{x} - x_{i-1}| \end{aligned}$$

For x_{i-1} "sufficiently close" to $\overline{x}$:

$$\begin{aligned} \frac{M_2}{2m_1} |\overline{x} - x_{i-1}| &\leq q \\ \frac{|\overline{x} - x_i|}{|\overline{x} - x_{i-1}|} &\leq q \end{aligned}$$

for some (fixed) $q < 1$, i.e., the error is reduced at least by a factor q in one iteration and the method converges.
Theorem: Let f have a continuous second derivative in a neighborhood of a **simple** root $\overline{x}$. Then Newton's method converges in some neighborhood of the root $\overline{x}$.

Alternative: Numerical approximation of the derivative as a part of the method.
We avoid the need of additional evaluations of the function by using the last two computed values, $f(x_{i-1})$, $f(x_{i-2})$; the derivative (=the slope of the tangent) is approximated by the slope of the secant:

$$\begin{aligned} f'(x_{i-1}) &\approx \frac{f(x_{i-1})-f(x_{i-2})}{x_{i-1}-x_{i-2}} \\ x_i &= x_{i-1}-\frac{f(x_{i-1})}{\frac{f(x_{i-1})-f(x_{i-2})}{x_{i-1}-x_{i-2}}} = \frac{x_{i-2}f(x_{i-1})-x_{i-1}f(x_{i-2})}{f(x_{i-1})-f(x_{i-2})} \end{aligned}$$

This is nothing new but the secant method (but the idea was good).
Definition: Let a method solving the equation $f(x)=0$ produce a sequence of approximations x_i, $i\in\mathbb{N}$, converging to a root $\overline{x}$. Then the **order of the method** is a number p such that the limit

$$\lim_{i\to\infty}\frac{|\overline{x}-x_i|}{|\overline{x}-x_{i-1}|^p}$$

exists and it is finite and nonzero.
For smaller p, $\frac{|\overline{x}-x_i|}{|\overline{x}-x_{i-1}|^p}\to 0$. For bigger p, $\frac{|\overline{x}-x_i|}{|\overline{x}-x_{i-1}|^p}\to\infty$. For at most one p the limit exists and it is finite and nonzero; this value is the order of the method.

method	order	condition
bisection	undef. (~ 1)	
regula falsi	1	the second derivative does not change its sign
secant	$(1+\sqrt{5})/2$	simple root
Newton's	2	simple root

Equation $f(x)=0$ can be tranformed to an equivalent form $\varphi(x)=x$,
e.g., by taking $\varphi(x)=f(x)+x$
Initial estimate x_0,

$$x_i=\varphi(x_{i-1})\,.$$

Ending condition:

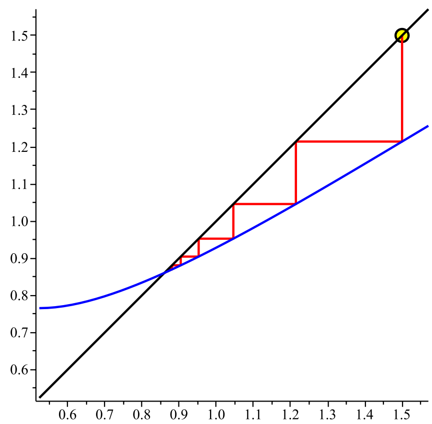
$$|x_i-x_{i-1}|<\eta\,.$$

Proposition: If the iteration method converges to $\tilde{x}$ where φ is continuous, then $\varphi(\tilde{x})=\tilde{x}$, $f(\tilde{x})=0$.

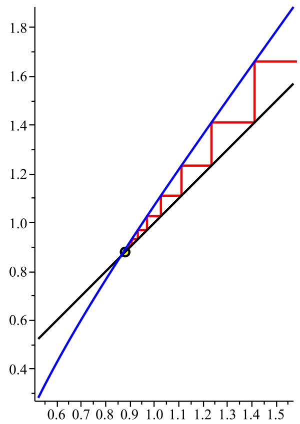
Proof:

$$\varphi(\tilde{x})=\varphi\Big(\lim_{i\to\infty}x_i\Big)=\lim_{i\to\infty}\varphi(x_i)=\lim_{i\to\infty}x_{i+1}=\lim_{i\to\infty}x_i=\tilde{x}\,.$$

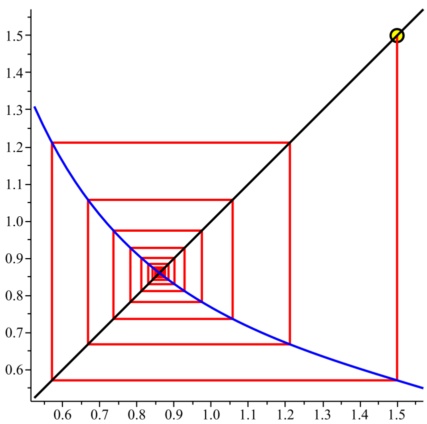
Find the least positive solution to the equation $f(x)=0$, where $f(x)=x-\cot x$.
$f(\frac{\pi}{4})=\frac{\pi}{4}-1<0,\quad f(\frac{\pi}{2})=\frac{\pi}{2}>0\quad\Rightarrow\quad\overline{x}\in(\frac{\pi}{4},\frac{\pi}{2})$
Choose $\varphi(x)=\lambda f(x)+x$, where $\lambda\neq 0$; ending condition for $\eta=0.001$.
We shall try $\lambda\in\{-0.2,0.2,-0.65,-0.8\}$.



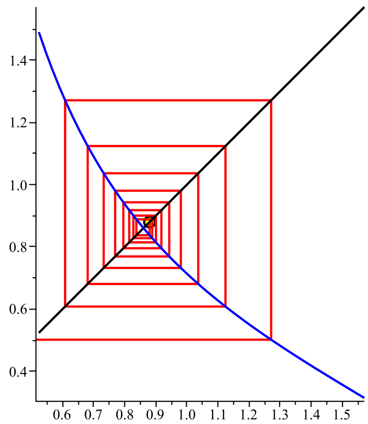
$x_{i+1}=0.8\,x_i+0.2\cot x_i$,
$x_0=1.5$,
converges monotonically



$x_{i+1}=1.2\,x_i-0.2\cot x_i$,
$x_0=0.88$
diverges monotonically



$x_{i+1} = 0.35\,x_i + 0.65\cot x_i$,
$x_0 = 1.5$,
converges non-monotonically



$x_{i+1} = 0.2\,x_i + 0.8\cot x_i$,
$x_0 = 0.88$,
diverges non-monotonically

Definition: A function φ is called **contractive** at interval I (with coefficient q) if

$$\exists q < 1\ \forall u, v \in I : |\varphi(u) - \varphi(v)| \le q \cdot |u - v|\,.$$

contractivity $\Rightarrow$ continuity

Theorem: (Sufficient condition for contractivity) Let function φ have a continuous derivative at interval I and let there exist $q < 1$ such that

$$\forall x \in I : |\varphi'(x)| \le q\,.$$

Then φ is contractive at I with coefficient q.

Proof:

$$|\varphi(u) - \varphi(v)| = \left|\int_v^u \varphi'(x)\,dx\right| \le \int_v^u |\varphi'(x)|\,dx \le \int_v^u q\,dx = q \cdot |u - v|\,.$$

Theorem: (Banach fixed point theorem for a real function) Let φ be a function contractive at a closed interval $I = [a, b]$ with coefficient $q < 1$ and let φ map I into I. Then the equation $\varphi(x) = x$ as a unique solution, $\overline{x}$, at interval I. This solution is the limit of the iteration method with an arbitrary initial value $x_0 \in I$ and the following estimate of the error holds:

$$|\overline{x} - x_i| \le \frac{q}{1-q}\,|x_i - x_{i-1}|\,.$$

Proof:

- Existence of the solution:

φ maps I into I

$\psi(x) = \varphi(x) - x$ is ψ non-negative at a and non-positive at b; it is continuous, hence it has a root in I, which is a solution to the equation $\varphi(x) = x$.

- Uniqueness of the solution: Assume that $\overline{\overline{x}} \in I$ is another solution. Then

$$|\overline{x} - \overline{\overline{x}}| = |\varphi(\overline{x}) - \varphi(\overline{\overline{x}})| \le q \cdot |\overline{x} - \overline{\overline{x}}| \quad \Rightarrow \quad \overline{\overline{x}} = \overline{x}$$

- Convergence of the iteration method to the solution:

$$|\overline{x} - x_i| = |\varphi(\overline{x}) - \varphi(x_{i-1})| \le q \cdot |\overline{x} - x_{i-1}| \le \ldots \le q^i \cdot |\overline{x} - x_0| \to 0$$

- Estimate of the error:

$$\begin{aligned} |\overline{x} - x_i| &\le q \cdot |\overline{x} - x_{i-1}| = q \cdot |(\overline{x} - x_i) + (x_i - x_{i-1})| \\ &\le q \cdot |\overline{x} - x_i| + q \cdot |x_i - x_{i-1}| \\ |\overline{x} - x_i| &\le \frac{q}{1-q}\,|x_i - x_{i-1}| \end{aligned}$$

How to transform the equations $f(x) = 0$ to an equivalent form $\varphi(x) = x$ such that the iteration method converges fast?
Possible solution:

$$\varphi(x) = x + \lambda f(x)\,,$$

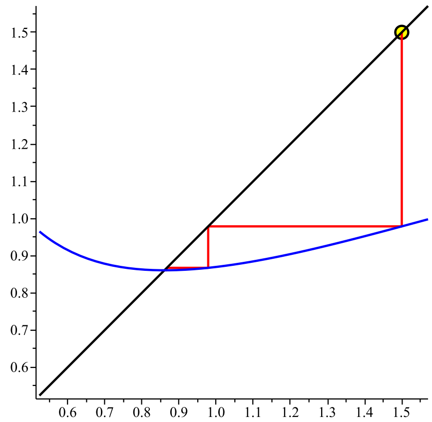
where $\lambda \neq 0$ and

$$\varphi'(x) = 1 + \lambda f'(x)$$

is small.

Example: (continued) $f'(x) = 2 + \cot^2 x \in [2,3] \Rightarrow \lambda \in [-\frac{1}{2}, -\frac{1}{3}]$
$-1/f'(0.86) \approx -0.365 \Rightarrow \lambda = -0.365$

$$\varphi(x) = 0.635\,x + 0.365\cot x\,.$$



$x_{i+1} = 0.635\,x_i + 0.365\cot x_i$, $x_0 = 1.5$
converges monotonically and fast

Theorem: Let the iteration method converge to $\overline{x}$. Let p be the least natural number such that $\varphi^{(p)}(\overline{x}) \neq 0$, and let $\varphi^{(p)}$ be continuous in some neighborhood of $\overline{x}$. Then the order of the iteration method is p.

Proof: Evaluate the Taylor series of function φ with center $\overline{x}$ at x_{i-1}:

$$\varphi(x_{i-1}) = \varphi(\overline{x}) + \frac{1}{p!}\,(x_{i-1} - \overline{x})^p\,\varphi^{(p)}(\xi_{i-1})\,, \text{ where } \xi_{i-1} \in I(\overline{x}, x_{i-1}),$$

$$x_i = \overline{x} + \frac{1}{p!}\,(x_{i-1} - \overline{x})^p\,\varphi^{(p)}(\xi_{i-1})\,,$$

$$\frac{|\overline{x} - x_i|}{|\overline{x} - x_{i-1}|^p} = \frac{1}{p!}\,|\varphi^{(p)}(\xi_{i-1})| \to \frac{1}{p!}\,|\varphi^{(p)}(\overline{x})| \in (0, +\infty)\,.$$

Comment: Usually $\varphi'(\overline{x}) \neq 0$, hence the iteration method is of order 1. However, this is not a rule, e.g., Newton's method is a special case of the iteration method.

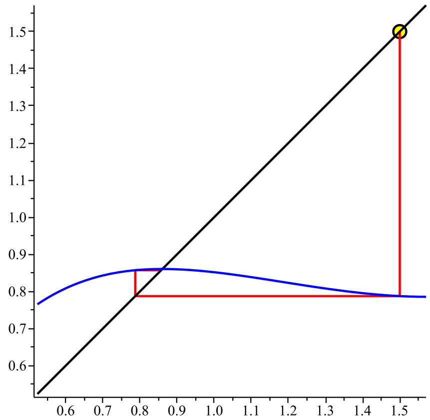
Idea: In each iteration, choose a new coefficient λ_i so that $\varphi'(x_i) = 0$, i.e.

$$\lambda_i = -\frac{1}{f'(x_i)}\,,$$

We get

$$x_{i+1} = x_i + \lambda_i\,f(x_i) = x_i - \frac{f(x_i)}{f'(x_i)}\,,$$

which is Newton's method (as a special case of the iteration method); this is usually of order 2, while the iteration method is usually of order 1.



$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$, $x_0 = 1.5$
converges non-monotonically and fast

- **single-point**, e.g., iteration method (which is in a sense a universal single-point method), Newton's method,
- **double-point**, e.g., bisection, regula falsi, secant method,
- **multiple-point**.

From the programmer's point of view:
- **not requiring the derivative**, e.g., bisection, regula falsi, secant method, and *usually* iteration method (depending on the iterative formula),
- **requiring the first derivative**, e.g., Newton's method,
- **requiring higher-order derivatives**.

According to convergence, we divide the methods to
- **always convergent**, e.g., bisection and regula falsi,
- **others**, e.g., Newton's, secant, and iteration method.
- In a neighborhood of a root of an *even multiplicity*, the function does not change its sign, hence bisection and regula falsi cannot be used.
- Secant and Newton's method are applicable to the computation of multiple roots, but in this case their convergence is only of the first order.
- In the iteration method, only the iterative formula, not the multiplicity of the root in the original equation, is important.

Method 1: Find all roots of function f' and test whether some of them is a root of function f. At roots of f of an even multiplicity, f' has a root of an odd multiplicity and changes its sign.
Method 2: Consider the function $h(x) = \frac{f(x)}{f'(x)}$ (where "removable discontinuities are removed").
Proposition: Let $\overline{x}$ be a root of function f of multiplicity k, such that in its neighborhood f has a continuous derivative of order k. Then $\overline{x}$ is a simple root of function $h = f/f'$.

Proof: The definition of a root of multiplicity k says that $f^{(j)}(\overline{x}) = 0$ if $j < k$ and $f^{(k)}(\overline{x}) \neq 0$. Repeated application of l'Hospital's rule results in a nonzero limit

$$\lim_{x\to\overline{x}} \frac{f(x)}{(x-\overline{x})^k} = \lim_{x\to\overline{x}} \frac{f'(x)}{k\,(x-\overline{x})^{k-1}} = \ldots = \lim_{x\to\overline{x}} \frac{f^{(k)}(x)}{k!} \neq 0\,.$$

Thus the proportion of the first two expressions is defined and equal to 1,

$$\lim_{x\to\overline{x}} \frac{f(x)}{(x-\overline{x})^k} \cdot \frac{k\,(x-\overline{x})^{k-1}}{f'(x)} = 1\,,$$

which implies the limit of h

$$\lim_{x\to\overline{x}} \frac{h(x)}{x-\overline{x}} = \frac{1}{k} \neq 0\,.$$

In the latter limit, the denominator converges to zero and so does the numerator, hence $\lim_{x\to\overline{x}} h(x) = 0$ and $\overline{x}$ is a root of function h. According to l'Hospital's rule, also $\lim_{x\to\overline{x}} h'(x)$ is nonzero, hence $\overline{x}$ is a simple root of function h.

If function f has only roots of finite multiplicities, then function h has the same roots, but simple (however, h is usually not continuous).
Special case of equation $f(x) = 0$, where f is a polynomial.
Theorem: (Estimate of the position of the roots of a polynomial) The absolute value of all (complex) roots of the equation

$$\sum_{i=0}^{n} a_i\, x^i = 0$$

is at most

$$1 + \frac{\max\bigl(|a_0|, \ldots, |a_{n-1}|\bigr)}{|a_n|}\,.$$

Bisection and regula falsi depend on the total order of reals $\Rightarrow$ not useful for higher dimensions.
The secant and Newton's method are applicable.
Complex roots can be found only for initial estimates with a nonzero imaginary part.
Newton's method can be generalized to systems of equations; the Jacobian plays the role of a derivative and instead of division, multiplication by an inverse matrix is needed. This causes higher complexity and problems with singularities. The conditions of convergence are more complex than in the one-dimensional case.
The iterative method is applicable; it may by difficult to maintain contractivity.
Because of problems with convergence, other methods were suggested which work on different principles not used in the one-dimensional case.