Asset and Risk Management: Risk Oriented Finance. Louis Esch, Robert Kieffer and Thierry Lopez.
© 2005 De Boeck & Larcier s.a. Published 2005 by John Wiley & Sons Ltd.

Appendix 8
Numerical Methods for Solving Nonlinear Equations[1]

An equation is said to be nonlinear when it involves terms of degree higher than 1 in the unknown quantity. These terms may be polynomial or capable of being broken down into Taylor series of degrees higher than 1.

Nonlinear equations cannot in general be solved analytically. In this case, therefore, the solutions of the equations must be approached using iterative methods. The principle of these methods of solving consists in starting from an arbitrary point – the closest possible point to the solution sought – and involves arriving at the solution gradually through successive tests.

The two criteria to take into account when choosing a method for solving nonlinear equations are:

- Method convergence (conditions of convergence, speed of convergence etc.).
- The cost of calculating of the method.

8.1 GENERAL PRINCIPLES FOR ITERATIVE METHODS

8.1.1 Convergence

Any nonlinear equation $f(x) = 0$ can be expressed as $x = g(x)$.

If x_0 constitutes the arbitrary starting point for the method, it will be seen that the solution x^* for this equation, $x^* = g(x^*)$, can be reached by the numerical sequence:

$$x_{n+1} = g(x_n) \quad n = 0, 1, 2, \ldots$$

This iteration is termed a Picard process and x^*, the limit of the sequence, is termed the fixed iterative point.

In order for the sequence set out below to tend towards the solution of the equation, it has to be guaranteed that this sequence will converge. A sufficient condition for convergence is supplied by the following theorem: if $x = g(x)$ has a solution a within the interval $I = [a - b; a + b] = \{x : |x - a| \leq b\}$ and if $g(x)$ satisfies Lipschitz's condition:

$$\exists \mathrm{L} \in [0; 1[\; : \; \forall x \in I, \quad |g(x) - g(a)| \leq L|x - a|$$

Then, for every $x_0 \in I$:

- all the iterated values x_n will belong to I;
- the iterated values x_n will converge towards a;
- the solution a will be unique within interval I.

[1] This appendix is mostly based on Litt F. X., *Analyse numérique, première partie*, ULG, 1999. Interested readers should also read: Burden R. L. and Faires D. J., *Numerical Analysis*, Prindle, Weber & Schmidt, 1981; and Nougier J. P., *Méthodes de calcul numérique*, Masson, 1993.

We should also show a case in which Lipschitz's condition is satisfied: it is sufficient that for every $x \in I$, $g'(x)$ exists and is such that $|g'(x)| \leq m$ with $m < 1$.

8.1.2 Order of convergence

It is important to choose the most suitable of the methods that converge. At this level, one of the most important criteria to take into account is the speed or order of convergence.

Thus the sequence x_n, defined above, and the error $e_n = x_n - a$. If there is a number p and a constant $C > 0$ so that

$$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^p} = C$$

p will then be termed the order of convergence for the sequence and C is the asymptotic error constant.

When the speed of convergence is unsatisfactory, it can be improved by the Aitken extrapolation,[2] which is a convergence acceleration process. The speed of convergence of this extrapolation is governed by the following result:

- If Picard's iterative method is of the order p, the Aitken extrapolation will be of the order $2p - 1$.
- If Picard's iterative method is of the first order, Aitken's extrapolation will be of the second order in the case of a simple solution and of the first order in the case of a multiple solution. In this last case, the asymptotic error constant is equal to $1 - 1/m$ where m is the multiplicity of the solution.

8.1.3 Stop criteria

As stated above, the iterative methods for solving nonlinear equations supply an approached solution to the solution of the equation. It is therefore essential to be able to estimate the error in the solution.

Working on the mean theorem:

$$f(x_n) = (x_n - a)f'(\xi), \text{ with } \xi \in [x_n; a]$$

we can deduce the following estimation for the error:

$$|x_n - a| \leq \frac{|f(x_n)|}{M}, \quad |f'(x_n)| \geq M, \quad x \in [x_n; a]$$

In addition, the rounding error inherent in every numerical method limits the accuracy of the iterative methods to:

$$\varepsilon_a = \frac{\delta}{f'(a)}$$

[2] We refer to Litt F. X., *Analyse numérique, première partie*, ULG 1999, for further details.

in which δ represents an upper boundary for the rounding error in iteration n:

$$\delta \geq |\delta_n| = \overline{f}(x_n) - f(x_n)$$

$\overline{f}(x_n)$ represents the calculated value for the function.

Let us now assume that we wish to determine a solution a with a degree of precision ε. We could stop the iterative process on the basis of the error estimation formula.

These formulae, however, require a certain level of information on the derivative $f'(x)$, information that is not easy to obtain. On the other hand, the limit specification ε_a will not generally be known beforehand.[3] Consequently, we are running the risk of ε, the accuracy level sought, never being reached, as it is better than the limit precision $\varepsilon_a(\varepsilon < \varepsilon_a)$. In this case, the iterative process will carry on indefinitely.

This leads us to accept the following stop criterion:

$$\begin{cases} |x_n - x_{n-1}| < \varepsilon \\ |x_{n+1} - x_n| \geq |x_n - x_{n-1}| \end{cases}$$

This means that the iteration process will be stopped when the iteration n produces a variation in value less than that of the iteration $n + 1$. The value of ε will be chosen in a way that prevents the iteration from stopping too soon.

8.2 PRINCIPAL METHODS

Defining an iterative method is based ultimately on defining the function $h(x)$ of the equation $x = g(x) \equiv x - h(x)f(x)$.

The choice of this function will determine the order of the method.

8.2.1 First order methods

The simplest choice consists of taking $h(x) = m = \text{constant} \neq 0$.

8.2.1.1 Chord method

This defines the *chord method* (Figure A8.1), for which the iteration is $x_{n+1} = x_n - mf(x_n)$.

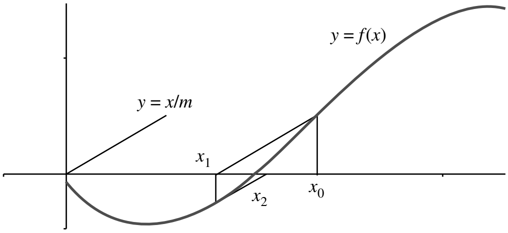

Figure A8.1 Chord method

[3] This will in effect require knowledge of $f'(a)$, when a is exactly what is being sought.

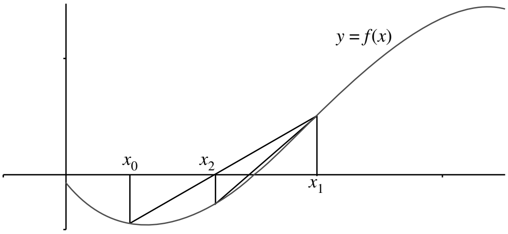

Figure A8.2 Classic chord method

The sufficient convergence condition (see Section A8.1.1) for this method is $0 < mf'(x) < 2$, in the neighbourhood of the solution. In addition, it can be shown that $\lim_{n\to\infty} \dfrac{|e_{n+1}|}{|e_n|} = |g'(a)| \neq 0$.

The chord method is therefore clearly a first-order method (see Section A8.1.2).

8.2.1.2 *Classic chord method*

It is possible to improve the order of convergence by making m change at each iteration:

$$x_{n+1} = x_n - m_n f(x_n)$$

The *classic chord method* (Figure A8.2) takes as the value for m_n the inverse of the slope for the straight line defined by the points $(x_{n-1}; f(x_{n-1}))$ and $(x_n; f(x_n))$:

$$x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n)$$

This method will converge if $f'(a) \neq 0$ and $f''(x)$ is continuous in the neighbourhood of a. In addition, it can be shown that

$$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^p} = \left(\frac{f''(a)}{2f'(a)}\right)^{1/p} \neq 0$$

for $p = \frac{1}{2}(1 + \sqrt{5}) = 1.618\ldots > 1$, which greatly improves the order of convergence for the method.

8.2.1.3 *Regula falsi method*

The *regula falsi* method (Figure A8.3) takes as the value for m_n the inverse of the slope for the straight line defined by the points $(x_{n'}; f(x_{n'}))$ and $(x_n; f(x_n))$ where n' is the highest index for which $f(x_{n'}).f(x_n) < 0$:

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

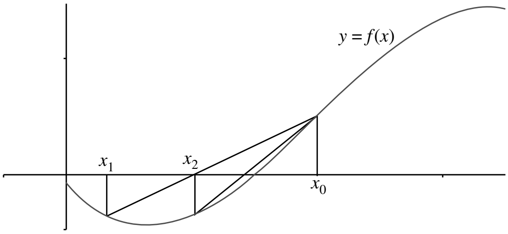

Figure A8.3 *Regula falsi* method

This method always converges when $f(x)$ is continuous. On the other hand, the convergence of this method is linear and therefore less effective than the convergence of the classic chord method.

8.2.2 Newton–Raphson method

If, in the classic chord method, we choose m_n so that $g'(x_n) = 0$, that is, $f'(x_n) = 1/m_n$, we will obtain a second-order iteration.

The method thus defined,

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

is known as the Newton-Raphson method (Figure A8.4).

It is clearly a second-order method, as

$$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^2} = \frac{1}{2}\left|\frac{f''(a)}{f'(a)}\right| \neq 0$$

The Newton–Raphson method is therefore rapid insofar as the initial iterated value is not too far from the solution sought, as global convergence is not assured at all.

A convergence criterion is therefore given for the following theorem. Assume that $f'(x) \neq 0$ and that $f''(x)$ does not change its sign within the interval $[a; b]$ and $f(a).f(b) < 0$.

If, furthermore,

$$\left|\frac{f(a)}{f'(a)}\right| < b - a \quad \text{and} \quad \left|\frac{f(b)}{f'(b)}\right| < b - a$$

the Newton–Raphson method will converge at every initial arbitrary point x_0 that belongs to $[a; b]$.

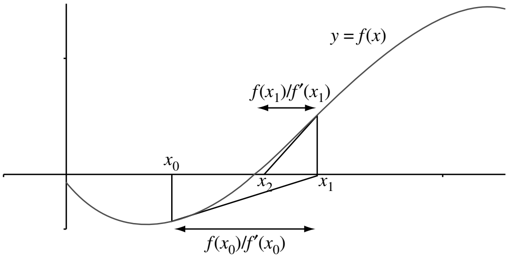

Figure 8.4 Newton–Raphson method

The classic chord method, unlike the Newton–Raphson method, requires two initial approximations but only involves one new function evaluation at each subsequent stage. The choice between the classic chord method and the Newton–Raphson method will therefore depend on the effort of calculation required for evaluation $f'(x)$.

Let us assume that the effort of calculation required for evaluation of $f'(x)$ is θ times the prior effort of calculation for $f(x)$.

Given what has been said above, we can establish that the effort of calculation will be the same for the two methods if:

$$\frac{1+\theta}{\log 2} = \frac{1}{\log p} \text{ in which } p = \frac{1+\sqrt{5}}{2}$$

is the order of convergence in the classic chord method.

In consequence:

- If $\theta > (\log 2/\log p) - 1 \sim 0.44 \rightarrow$ the classic chord method will be used.
- If $\theta \leq (\log 2/\log p) - 1 \sim 0.44 \rightarrow$ the Newton–Raphson method will be used.

8.2.3 Bisection method

The *bisection method* is a linear convergence method and is therefore slow. Use of the method is, however, justified by the fact that it converges overall, unlike the usual methods (especially the Newton–Raphson and classic chord methods). This method will therefore be used to bring the initial iterated value of the Newton–Raphson or classic chord method to a point sufficiently close to the solution to ensure that the methods in question converge.

Let us assume therefore that $f(x)$ is continuous in the interval $[a_0; b_0]$ and such that[4] $f(a_0).f(b_0) < 0$. The principle of the method consists of putting together a converging sequence of bracketed intervals, $[a_1; b_1] \supset [a_2; b_2] \supset [a_3; b_3] \supset \ldots$, all of which contain a solution of the equation $f(x) = 0$.

If it is assumed that[5] $f(a_0) < 0$ and $f(b_0) > 0$, the intervals $I_k = [a_k; b_k]$ will be put together by recurrence on the basis of I_{k-1}.

$$[a_k; b_k] = \begin{cases} [m_k; b_{k-1}] & \text{if} \quad f(m_k) < 0 \\ [a_{k-1}; m_k] & \text{if} \quad f(m_k) > 0 \end{cases}$$

Here, $m_k = (a_{k-1} + b_{k-1})/2$. One is thus assured that $f(a_k) < 0$ and $f(b_k) > 0$, which guarantees convergence.

The bisection method is not a Picard iteration, but the order of convergence can be determined, as $\lim_{n\to\infty} \dfrac{|e_{n+1}|}{|e_n|} = \dfrac{1}{2}$. The bisection method is therefore a first-order method.

8.3 NONLINEAR EQUATION SYSTEMS

We have a system of n nonlinear equations of n unknowns: $f_i(x_1, x_2, \ldots, x_n) = 0 \quad i = 1, 2, \ldots, n$. Here, in vectorial notation, $f(x) = 0$. The solution to the system is an n-dimensional vector a.

[4] This implies that $f(x)$ has a root within this interval.

[5] This is not restrictive in any way, as it corresponds to $f(x) = 0$ or $-f(x) = 0, x \in [a_0; b_0]$, depending on the case.

8.3.1 General theory of n-dimensional iteration

n-dimensional iteration general theory is similar to the one-dimensional theory. The above equation can thus be expressed in the form:

$$x = g(x) \equiv x - \mathbf{A}(x)f(x)$$

where $\mathbf{A}$ is a square matrix of n^{th} order.

Picard's iteration is always defined as

$$x_{k+1} = g(x_k) \quad k = 0, 1, 2 \text{ etc.}$$

and the convergence theorem for Picard's iteration remains valid in n dimensions.

In addition, if the Jacobian matrix $\mathbf{J}(x)$, defined by $[\mathbf{J}(x)]_{ij} = \big(g_j(x)\big)'_{x_i}$ is such that for every $x \in I$, $\|\mathbf{J}(x)\| \leq m$ for a norm compatible with $m < 1$, Lipschitz's condition is satisfied.

The order of convergence is defined by

$$\lim_{k\to\infty} \frac{\|e_{k+1}\|}{\|e_k\|^p} = C$$

where C is the constant for the asymptotic error.

8.3.2 Principal methods

If one chooses a constant matrix $\mathbf{A}$ as the value for $\mathbf{A}(x)$, the iterative process is the generalisation in n dimensions of the chord method.

If the inverse of the Jacobian matrix of f is chosen as the value of $\mathbf{A}(x)$, we will obtain the generalisation in n dimensions of the Newton–Raphson method.

Another approach to solving the equation $f(x) = 0$ involves using the i^{th} equation to determine the $(i + 1)^{\text{th}}$ component. Therefore, for $i = 1, 2, \ldots, n$, the following equations will be solved in succession:

$$f_i(x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}, x_i, x_{i+1}^{(k)}, \ldots, x_n^{(k)}) = 0$$

with respect to x_i. This is known as the *nonlinear Gauss–Seidel method.*