3 Numerical Solution of Nonlinear Equations and Systems

3.1 Fixed point iteration

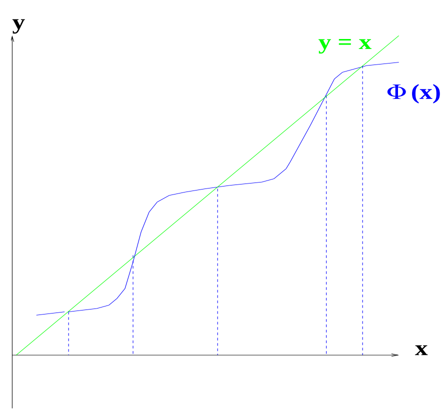
Reamrk 3.1 Problem

Given a function $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^n$, compute $\mathbf{x}^* \in \mathbb{R}^n$ such that

$$(*) \qquad \mathbf{F}(\mathbf{x}^*) = \mathbf{0} .$$

In this chapter, we consider the iterative solution of $(*)$.

Definition 3.2 Fixed point, fixed point iteration



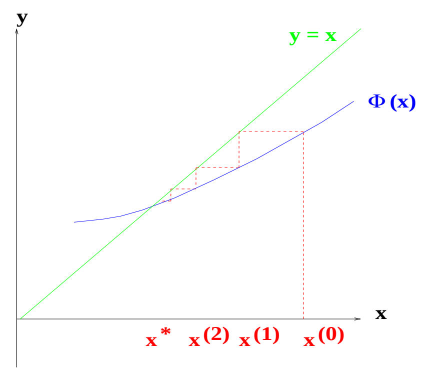
Let $\mathbf{\Phi} : \mathbb{R}^n \to \mathbb{R}^n$ be given. A vector $\mathbf{x}^* \in \mathbb{R}^n$ is called a fixed point of $\mathbf{\Phi}$, if

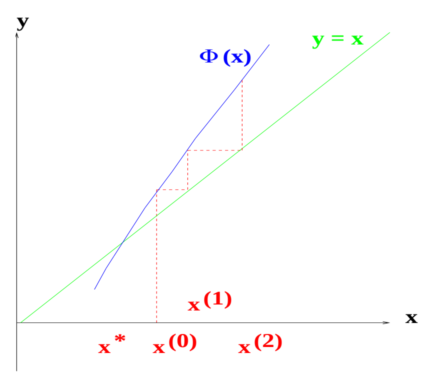
$$\mathbf{\Phi}(\mathbf{x}^*) = \mathbf{x}^* .$$

Moreover, given a start vector $\mathbf{x}^{(0)} \in \mathbb{R}^n$, the iteration

$$\mathbf{x}^{(k+1)} = \mathbf{\Phi}(\mathbf{x}^{(k)}) \quad , \quad k \in \mathbb{N}_0$$

is said to be a fixed point iteration (method of successive approximations).





Remark 3.3 Attractive and repulsive fixed points

The two figures show that a fixed point iteration may be convergent as well as divergent.
In these cases, the fixed point is called attractive resp. repulsive.

Definition 3.4 Contraction

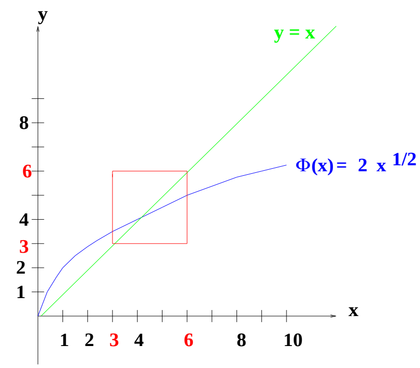
A mapping $\Phi : \bar{D} \subset \mathbb{R}^n \to \mathbb{R}^n$ is said to be a contraction on $\bar{D}$, if there exists $0 \leq \kappa < 1$ such that

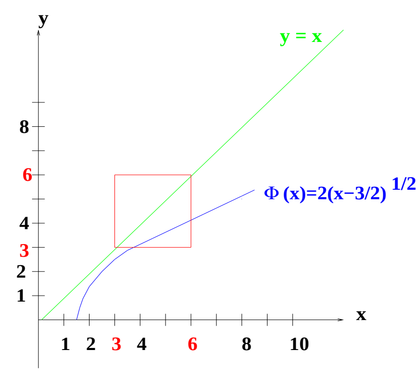
$$\|\Phi(x_1) - \Phi(x_2)\| \leq \kappa \, \|x_1 - x_2\| \quad , \quad x_1, x_2 \in \bar{D} \ .$$

The number κ is referred to as the contraction number.

Remark 3.5 Sufficient conditions for a contraction

In case $n = 1$ and $\bar{D} = [a,b] \subset \mathbb{R}$ assume $\Phi \in C^1([a,b])$ with $\kappa := \max_{z \in [a,b]} |\Phi'(z)| < 1$. Then, Φ is a contraction on $[a,b]$ with contraction number κ.





Remark 3.5 Existence of fixed points

In general, the property to be a contraction is not sufficient for the existence of fixed points:

(i) The mapping $\Phi(x) = 2\sqrt{x}$ has a fixed poinr in $x = 4$.

We have $|\Phi'(x)| < 1\ ,\ x \in [3,6]$.

Moreover, we have $\Phi(3) > 3\ ,\ \Phi(6) < 6 \quad \Longrightarrow \quad \Phi([3,6]) \subset [3,6]$.

(ii) The mapping $\Phi(x) = 2\sqrt{x - 3/2}$ does not have a fixed point.

Here, we also have $|\Phi'(x)| < 1\ ,\ x \in [3,6]$.

On the other hand, we have $\Phi(3) < 3 \quad \Longrightarrow \quad \Phi([3,6]) \not\subset [3,6]$.

Theorem 3.7 The Banach fixed point theorem

Let $D \subset \mathbb{R}^n$ and $\Phi : \bar{D} \to \mathbb{R}^n$ a mapping such that

$$(1) \quad \Phi(\bar{D}) \subset \bar{D} \, ,$$

$$(2) \quad \|\Phi(x_1) - \Phi(x_2)\| \leq \kappa \, \|x_1 - x_2\| \quad , \quad x_1, x_2 \in \bar{D} \, , \; 0 \leq \kappa < 1 \, .$$

Then, there holds:

(i) The mapping Φ has a unique fixed point $x^* \in \bar{D}$.

(ii) For any start vector $x^{(0)} \in \bar{D}$, the fixed point iteration converges satisfying the a priori estimate

$$(*) \quad \|x^{(k)} - x^*\| \leq \frac{\kappa^k}{1 - \kappa} \|x^{(1)} - x^{(0)}\| \quad , \quad k \in \mathbb{N} \, ,$$

and the a posteriori estimate

$$(**) \quad \|x^{(k)} - x^*\| \leq \frac{\kappa}{1 - \kappa} \|x^{(k)} - x^{(k-1)}\| \quad , \quad k \in \mathbb{N} \, .$$

Proof: We have:

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| \le \|\Phi(\mathbf{x}^{(k)} - \Phi(\mathbf{x}^{(k-1)}\| \le \kappa \, \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \le \dots \le \kappa^k \, \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| \, .$$

We prove that $(\mathbf{x}^{(k)})_{k\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{R}^n$:

$$\begin{aligned}(+) \quad \|\mathbf{x}^{(k+m)} - \mathbf{x}^{(m)}\| &\le \|\mathbf{x}^{(k+m)} - \mathbf{x}^{(k+m-1)}\| + \dots + \|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| \le \\ &\le (\kappa^{k+m-1} + \dots \kappa^k) \, \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| = \\ &= \kappa^k \, (\kappa^{m-1} + \dots + 1) \, \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| \le \frac{\kappa^k}{1-\kappa} \, \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| \, .\end{aligned}$$

Hence, there exists $\mathbf{x}^* \in \bar{D}$ such that $\mathbf{x}^{(k)} \to \mathbf{x}^* \ (k \to \infty)$. Since $\mathbf{x}^{(k)} \in \bar{D}$, $k \in \mathbb{N}$, and $\bar{D}$ is closed, we conclude $\mathbf{x}^* \in \bar{D}$.

$\mathbf{x}^*$ is a fixed poibt of Φ, since

$$\begin{aligned}\|\mathbf{x}^* - \Phi(\mathbf{x}^*)\| &= \|\mathbf{x}^* - \mathbf{x}^{(k+1)} + \mathbf{x}^{(k+1)} - \Phi(\mathbf{x}^*)\| = \|\mathbf{x}^* - \mathbf{x}^{(k+1)} + \Phi(\mathbf{x}^{(k)}) - \Phi(\mathbf{x}^*)\| \text{ le} \\ &\le \|\mathbf{x}^* - \mathbf{x}^{(k+1)}\| + \kappa\|\mathbf{x}^{(k)} - \mathbf{x}^*\| \to 0 \quad (k \to \infty) \, .\end{aligned}$$

Proof of uniqueness:

Assume that $x_1^*, x_2^* \in \bar{D}$ are two fixed points of Φ. Then, there holds:

$$0 \leq \|x_1^* - x_2^*\| = \|\Phi(x_1^*) - \Phi(x_2^*)\| \leq \kappa \|x_1^* - x_2^*\| .$$

Hence, observing $\kappa < 1$, it follows that $x_1^* = x_2^*$.

The a priori estimate $(*)$ can be deduced from $(+)$ by passing to the limit $k \to \infty$.

The a posteriori estimate $(**)$ follows from $(*)$, if we choose $x^{(k-1)}$ as a startvector.

Remark 3.8 The importance of the Banach fixed point theorem
Theorem 3.7 is a special case of the Banach fixed point theorem which more generally holds true in complete metric spaces.

Remark 3.9 A priori and a posteriori estimates
Given an accuracy $0 < \varepsilon \ll 1$, before initiating the iteration, the a priori estimate $(*)$ allows to give an upper bound of the number of iterations that are necessary to obtain the prescribed accuracy. After an actual iteration, the a posteriori estimate $(**)$ can be used as a termination criterion.

Definition 3.10 Order of convergence

A sequence of iterates $(x^{(k)})_{k\in\mathbb{N}}$, $x^{(k)} \in \mathbb{R}^n$, $k \in \mathbb{N}$, converges to $x^* \in \mathbb{R}^n$ of order $p \ge 1$, if there exists a constant $C \ge 0$ such that:

$$\|x^{(k+1)} - x^*\| \le C \, \|x^{(k)} - x^*\|^p \quad , \quad k \in \mathbb{N} \, ,$$

and if p is the largest number for which such an estimate holds true.
In case $p = 1$, we have to require $C < 1$.
If $p = 1$, the convergence is said to be linear; if $p = 2$, it is called quadratic, etc.
The sequence is called superlinearly convergent, if there exists a null sequence $(C_k)_{k\in\mathbb{N}}$ of nonnegative numbers C_k , $k \in \mathbb{N}_0$, such that

$$\|x^{(k+1)} - x^*\| \le C_k \, \|x^{(k)} - x^*\| \quad , \quad k \in \mathbb{N}_0 \, .$$

Lemma 3.11 Order of convergence in case $n = 1$

Assume that $\Phi \in C^p([a,b])$, $[a,b] \subset \mathbb{R}$, and that $x^{(}*) \in [a,b]$ is a fixed point of Φ satisfying $\Phi^{(i)}(x^*) = 0, 1 \le i \le p-1$, and $\Phi^{(p)}(x^*) \ne 0$. Then, the fixed point iteration converges of order p.

Proof: Taylor expansion yields

$$x^{(k+1)} - x^* = \Phi(x^{(k)}) - \Phi(x^*) = \frac{1}{p!} \, \Phi^{(p)}(\xi) \, (x^{(k)} - x^*)^p \quad , \quad \xi \in [\min(x^{(k)}, x^*), \max(x^{(k)}, x^*)] \, .$$

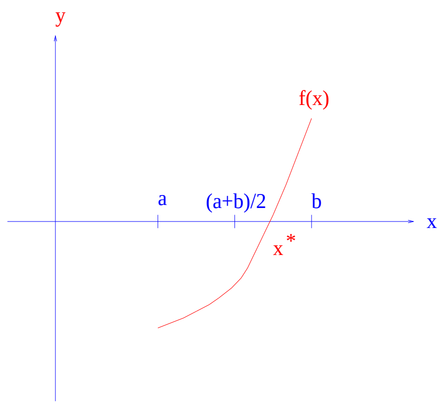
3.2 Special methods for scalar nonlinear equations

3.2.1 Method of bisection

Let us assume that there exist points $a < b$ satisfying

$$f(a)\, f(b) \ < \ 0 \, .$$

Further, suppose that $f \in C([a,b])$. Then, in (a,b) the function f has at least one zero x^*. The interval $[a,b]$ is said to be an inclusion interval.



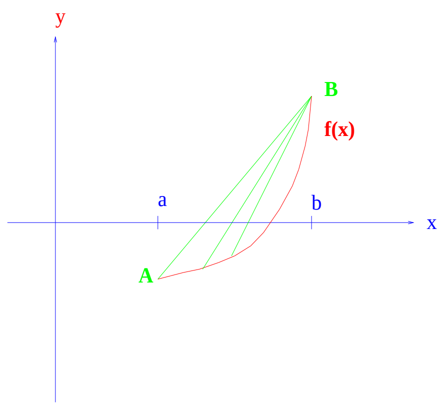
If we consider the sign of $f((a+b)/2)$, we obtain a new inclusion interval $[a,(a+b)/2]$ or $[(a+b)/2,b]$.
Given a tolerance $\varepsilon > 0$, a continuation of this approach yields:

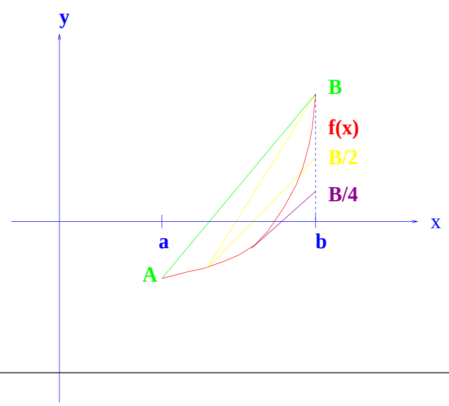
Method of bisection:

$A \ := \ f(a) \ ; \ B \ := \ f(b)$ with $A \cdot B < 0$;
as long as $b \ - \ a > \varepsilon$ compute:

$\quad t := (a+b)/2 \ ; \ T := f(t)$;
$\quad$ if $\ A \cdot T > 0 \ : \ a := t \ ; \ A := T$;
$\quad$ otherwise $\quad b := t \ ; \ B := T$;

$t := (a+b)/2$.





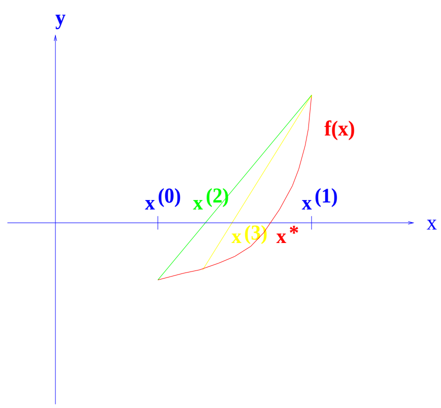
3.2.2 Regula falsi and the Illinois algorithm

The differerence between the regula falsi and the method of bisection is in the choice of t: t is chosen as the zero of the secant joining the points (a, A) , (b, B):

$$t = a + \frac{A}{A - B}(b - a) \quad .$$

The method converges linearly.

The Illinois algorithm realizes that t does not stay on the same side of the zero x^*:
After each evaluation of the function (except the first one), a function value will be halved.
A convergence analysis, taking into account various possible cases, results in the convergence order $p \sim 1.442$.



3.2.3 The secant method

Given two initial points $x^{(0)}$, $x^{(1)}$, the new iterate $x^{(2)}$ is computed as the zero of the secant joining the points $(x^{(0)}, f(x^{(0)}))$ and $(x^{(1)}, f(x^{(1)}))$:

$$S_{0,1}(x) \;=\; f(x^{(1)}) \;+\; \frac{f(x^{(1)}) \;-\; f(x^{(0)})}{x^{(1)} \;-\; x^{(0)}}\,(x \;-\; x^{(1)}) \quad .$$

We obtain:

$$x^{(2)} \;=\; x^{(1)} \;-\; \frac{x^{(1)} \;-\; x^{(0)}}{f(x^{(1)}) \;-\; f(x^{(0)})}\, f(x^{(1)}) \quad .$$

This leads to the iteration:

$$x^{(k+1)} \;=\; x^{(k)} \;-\; \frac{x^{(k)} \;-\; x^{(k-1)}}{f(x^{(k)}) \;-\; f(x^{(k-1)})}\, f(x^{(k)}) \quad , \quad k \in \mathbb{N} \; .$$

Hence, the secant method is a three-term recursion.

Remark 3.12 Order of convergence of the secant method

The secant

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

corresponds to the linear polynomial interpolating the function f in the points $(x^{(k-1)}, f(x^{(k-1)}))$ and $(x^{(k)}, f(x^{(k)}))$. Consequently, assuming $f \in C^2(\mathbb{R})$, for some $z \in \mathbb{R}$ (cf. Chapter 4):

$$f(z) - S_{k-1,k}(z) = \frac{1}{2} f''(\xi) (z - x^{(k-1)}) (z - x^{(k)}) \quad , \quad \xi \in [\min(x^{(k-1)}, x^{(k)}, z), \max(x^{(k-1)}, x^{(k)}, z)] \, .$$

Choosing $z = x^*$ (x^* zero of f) and observing (expansion of $S_{k-1,k}(x^*)$ around $x^{(k+1)}$:

$$S_{k-1,k}(x^*) = \underbrace{S_{k-1,k}(x^{(k+1)})}_{=0} + \frac{f(x^{(k)}) - f(x^{(k-1)})}{x^{(k)} - x^{(k-1)}} (x^* - x^{(k+1)}) \, ,$$

it follows that:

$$\frac{f(x^{(k)}) - f(x^{(k-1)})}{x^{(k)} - x^{(k-1)}} (x^{(k+1)} - x^*) = \frac{1}{2} f''(\xi) (x^* - x^{(k-1)}) (x^* - x^{(k)}) \, .$$

We remind:

$$\frac{f(x^{(k)}) - f(x^{(k-1)})}{x^{(k)} - x^{(k-1)}} (x^{(k+1)} - x^*) = \frac{1}{2} f''(\xi) (x^* - x^{(k-1)}) (x^* - x^{(k)}) .$$

The mean value theorem implies the existence of $\eta \in [\min(x^{(k-1)}, x^{(k)}), \max(x^{(k-1)}, x^{(k)})]$ satisfying

$$\frac{f(x^{(k)}) - f(x^{(k-1)})}{x^{(k)} - x^{(k-1)}} = f'(\eta) ,$$

such that for $f'(\eta) \neq 0$:

$$x^{(k+1)} - x^* = \frac{f''(\xi)}{2f'(\eta)} (x^{(k-1)} - x^*) (x^{(k)} - x^*) .$$

If x^* is a simple zero of f, i.e., $f'(x^*) \neq 0$, then there is an interval $I_\rho(x^*) = \{x \in \mathbb{R} \mid |x - x^*| \leq \rho\}$ such that $f'(x) \neq 0$, $x \in I_\rho(x^*)$, and we have:

$$|x^{(k+1)} - x^*| \leq c \, |x^{(k)} - x^*| \, |x^{(k-1)} - x^*| \quad , \quad c := (\max_{x \in I_\rho(x^*)} |f''(x)|)/(2 \min_{x \in I_\rho(x^*)} |f'(x)|) .$$

Multiplying

$$|x^{(k+1)} - x^*| \le c\,|x^{(k)} - x^*|\,|x^{(k-1)} - x^*| \quad , \quad c := \big(\max_{x \in I_\rho(x^*)} |f''(x)|\big)/\big(2 \min_{x \in I_\rho(x^*)} |f'(x)|\big) .$$

by c and computing the logarithm, yields:

$$\log\left(c\,|x^{(k+1)} - x^*|\right) \le \log\left(c\,|x^{(k)} - x^*|\right) + \log\left(c\,|x^{(k-1)} - x^*|\right) .$$

Setting $\varepsilon_k := \log\left(c\,|x^{(k)} - x^*|\right)$, we obtain the Fibonacci recursion

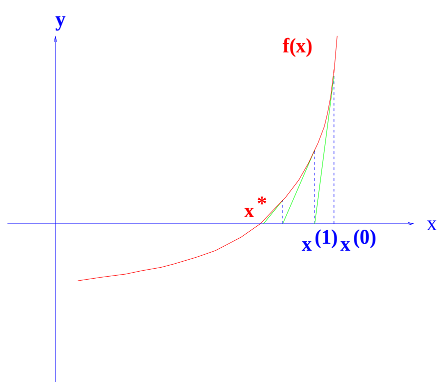
$$\varepsilon_{k+1} = \varepsilon_k + \varepsilon_{k-1} ,$$

whose solution is given by $\varepsilon_k = \alpha\,\lambda_1^k + \beta\,\lambda_2^k \quad , \quad \lambda_{1,2} = (1 \pm \sqrt{5})/2$.

For $k \to \infty$, we have $\varepsilon_k \sim \alpha\lambda_1^k$, and hence, $\varepsilon_{k+1} \sim \lambda_1 \varepsilon_k$, whence:

$$\log\left(c\,|x^{(k+1)} - x^*|\right) \le \log\left(c^{\lambda_1}\,|x^{(k)} - x^*|^{\lambda_1}\right) \quad \Longrightarrow \quad |x^{(k+1)} - x^*| \le c^{\lambda_1 - 1}\,|x^{(k)} - x^*|^{\lambda_1} .$$

The secant method is therefore convergent of order $p = (1 + \sqrt{5})/2 \sim 1.618$.



3.2.4 Newton's method

Given an initial value $x^{(0)}$, compute a new iterate $x^{(1)}$ as the zero of the tangent at the point $(x^{(0)}, f(x^{(0)}))$:

$$T_0(x) \ = \ f(x^{(0)}) \ + \ f'(x^{(0)} \ (x \ - \ x^{(0)}) \ .$$

We obtain:

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

This leads to the iteration:

$$x^{(k+1)} \ = \ x^{(k)} \ - \ \frac{f(x^{(k)})}{f'(x^{(k)})} \quad , \quad k \in \mathbb{N} \ .$$

Theorem 3.13 Convergence of Newton's method in 1D

Assume that $f \in C^2(\mathbb{R})$ and thatx^* is a simple zero, i.e., $f'(x^*) \neq 0$. Then, Newton's method is of local quadratic convergence.

Proof: The Newton iteration is equivalent to the fixed point iteration with respect to the function

$$\Phi(x) \; = \; x \; - \; \frac{f(x)}{f'(x)} \quad \Longrightarrow \quad \Phi'(x) \; = \; \frac{f(x)\, f''(x)}{(f'(x))^2} \; .$$

By assumption, there exists an interval $I_\rho(x^*) := \{x \in \mathbb{R} | |x - x^*| \leq \rho\}$, such that $|\Phi'(x)| < 1 \; , \; x \in I_\rho(x^*)$,

which implies convergence for initial values $x^{(0)} \in I_\rho(x^*)$ (local convergence).

Moreover, by Taylor expansion we find:

$$f(x) \; = \; f(x^{(k)}) \; + \; f'(x^{(k)}) \; (x \; - \; x^{(k)}) \; + \; \frac{1}{2} \, f''(\xi) \; (x \; - \; x^{(k)})^2 \quad , \quad \xi \in [\min(x, x^{(k)}), \max(x, x^{(k)})] \; .$$

Choosing $x = x^*$ and observing $f(x^{(k)}) = f'(x^{(k)})(x^{(k)} - x^{(k+1)})$, it follows that:

$$x^{(k+1)} \; - \; x^* \; = \; \frac{1}{2} \, \frac{f''(\xi)}{f'(x^{(k)})} \, (x^{(k)} \; - \; x^*)^2 \; .$$

Corollary 3.14 Convergence of Newton's method in case of multiple zeroes

Assume that $f \in C^p(\mathbb{R})$ and let x^* be a zero of multiplicity $p \geq 2$, i.e., $f^{(j)}(x^*) = 0 \ , \ 1 \leq j \leq p-1$, and $f^{(p)}(x^*) \neq 0$. Then, Newton's method is locally linearly convergent.

Using the modified iteration

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

the method converges locally quadratically.

Proof: There exists a differentiable function g with $g(x^* \neq 0$ such that

$$\begin{aligned} f(x) &= (x - x^*)^p \, g(x) \ , \\ f'(x) &= p \, (x - x^*)^{p-1} \, g(x) + (x - x^*)^p \, g'(x) \ . \end{aligned}$$

Hence, we obtain

$$\begin{aligned} \Phi(x) &= x - \frac{(x - x^*) \, g(x)}{p \, g(x) + (x - x^*) \, g'(x)} \ , \\ \Phi'(x^*) &= 1 - \tfrac{1}{p} \ . \end{aligned}$$

For the modified method we have $\Phi'(x^*) = 0$.

3.3 Newton's method for nonlinear systems

Remark 3.15 Setting the problem:

We consider the application of Newton's method to the nonlinear system

$$(*) \quad F(x) = 0 ,$$

where we assume $F : \mathbb{R}^n \to \mathbb{R}^n$ to be continuously differentiable. If $F = (F_1, ..., F_n)$, then $(*)$ can be written in more detail as follows:

$$\begin{aligned} F_1(x_1, ..., x_n) &= 0 , \\ F_2(x_1, ..., x_n) &= 0 , \\ \cdot \qquad &\quad \cdot , \\ F_n(x_1, ..., x_n) &= 0 . \end{aligned}$$

Given an initial vector $x^{(0)} \in \mathbb{R}^n$, we locally approximate F by

$$F(x) = F(x^{(0)}) + F'(x^{(0)}) (x - x^{(0)}) + o(\|x - x^{(0)}\|) .$$

where $F'(x^{(0)})$ denotes the Jacobi matrix in $x^{(0)}$.

$$F(x) = F(x^{(0)}) + F'(x^{(0)})\,(x - x^{(0)}) + o(\|x - x^{(0)}\|)\,.$$

The Jacobi matrix $F'(x^{(0)})$ is the matrix of the partial derivatives of F in $x^{(0)}$:

$$F'(x^{(0)}) = \begin{pmatrix} \frac{\partial F_1}{\partial x_1}(x^{(0)}) & \cdots & \frac{\partial F_1}{\partial x_n}(x^{(0)}) \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ \frac{\partial F_n}{\partial x_1}(x^{(0)}) & \cdots & \frac{\partial F_n}{\partial x_n}(x^{(0)}) \end{pmatrix}$$

We compute the new iterate $x^{(1)}$ as the zero of the linear approximation $\bar{F}(x) = F(x^{(0)}) + F'(x^{(0)})(x - x^{(0)})$, i.e., as the solution of the linear algebraic system

$$F'(x^{(0)})\,(x^{(1)} - x^{(0)}) = -\,F(x^{(0)})\,.$$

This gives rise to Newton's method

$$(**) \qquad \begin{cases} F'(x^{(k)})\,\Delta x^{(k)} = -\,F(x^{(k)})\,, \\ x^{(k+1)} = x^{(k)} + \Delta x^{(k)} \end{cases} .$$

The vectors $\Delta x^{(k)}$ are called the Newton increments.

Remark 3.16 Affine invariance of Newton's method

Let $A \in \mathbb{R}^{n \times n}$ be regular. Obviously, $x^* \in \mathbb{R}^n$ is a solution of $F(x) = 0$ if and only if x^* also solves the linear system

$$G(x) \; := \; A\,F(x) \; = \; 0 \; .$$

Hence, the nonlinear system $(*)$ is invariant with respect to affine transformations (without translations) in the range space. The same property holds true for Newton's method, since

$$(G'(x))^{-1}G(x) \; = \; (F'(x))^{-1}A^{-1}AF(x) \; = \; (F'(x))^{-1}F(x) \; .$$

Consequently, Newton's method is said to be affine invariant. For the same reason, the convergence properties of Newton's method should be investigated within the framework of an affine invariant convergence theory.

Theorem 3.17 Convergence of Newton's method in $\mathbb{R}^n$

Assume that $D \subset \mathbb{R}^n$ is an open, convex set and $F \in C^1(D)$ with regular Jacobi matrix $F'(x)\ ,\ x \in D$. Moreover, suppose that the following assumptions hold true:

(1) There exists a number $\omega \geq 0$, such that for all $\lambda \in [0,1]\ ,\ x \in D$ and $y \in \mathbb{R}^n$ with $x + \lambda y \in D$:

$$(+) \quad \|(F'(x))^{-1}\,[\,F'(x + \lambda y) - F'(x)\,]\,y\| \leq \lambda\,\omega\,\|y\|^2$$

(affine invariant Lipschitz condition for the Jacobi matrix)

(ii) There exist a solution $x^* \in D$ of $(*)$ and a start vector $x^{(0)} \in D$ such that

$$(++) \quad \rho := \|x^{(0)} - x^*\| < \frac{2}{\omega}\ ,$$

(iii) There holds: $\quad (+++) \quad K_\rho(x^*) := \{\, x \in \mathbb{R}^n \mid \|x - x^*\| < \rho \,\} \subset D$.

If $(x^{(k)})_{k \in \mathbb{N}}$ is the sequence generated by Newton's method, we have:

$$x^{(k)} \in K_\rho(x^*)\ ,\ k \in \mathbb{N} \quad \text{and} \quad \lim_{k \to \infty} x^{(k)} = x^*\ .$$

The solution is unique in $K_{2/\omega}(x^*)$, and there holds:

$$\|x^{(k+1)} - x^*\| \leq \frac{\omega}{2}\,\|x^{(k)} - x^*\|^2 \quad , \quad k \in \mathbb{N}_0\ .$$

Proof: We have:

$$(\circ)\quad x^{(k+1)} - x^* = x^{(k)} - (F'(x^{(k)}))^{-1} F(x^{(k)}) - x^* = x^{(k)} - x^* - (F'(x^{(k)}))^{-1} [\, F(x^{(k)}) - F(x^*) \,] =$$
$$= (F'(x^{(k)}))^{-1} [\, F(x^*) - F(x^{(k)}) - F'(x^{(k)})(x^* - x^{(k)}) \,] \, .$$

Using the mean value theorem of calculus, for $x, y \in D$ there holds:

$$F(y) - F(x) = \int_0^1 F'(x + \lambda\, (y - x))\, (y - x)\, d\lambda \, ,$$

and hence,

$$F(y) - F(x) - F'(x)\, (y - x) = \int_0^1 [\, F'(x + \lambda\, (y - x)) - F'(x) \,]\, (y - x)\, d\lambda \, ,$$

whence

$$\|(F'(x))^{-1} [F(y) - F(x) - F'(x)\, (y - x)]\| \le \int_0^1 \|(F'(x))^{-1} [F'(x + \lambda\, (y - x)) - F'(x)]\, (y - x)\|\, d\lambda \, .$$

$$\|(F'(x))^{-1}\,[F(y) - F(x) - F'(x)\,(y - x)]\| \le \int_0^1 \|(F'(x))^{-1}\,[F'(x + \lambda\,(y - x)) - F'(x)]\,(y - x)\|\,d\lambda\,.$$

Due to the convexity of D we have $x + \lambda(y - x) \in D$, and hence, assumption $(+)$ implies:

$$\|(F'(x))^{-1}\,[F(y) - F(x) - F'(x)\,(y - x)]\| \le \int_0^1 \lambda\,\omega\,\|y - x\|^2\,d\lambda = \frac{\omega}{2}\,\|y - x\|^2\,.$$

Setting $y = x^*$ and $x = x^{(k)}$, we deduce from $(\circ)$ that:

$$\|x^{(k+1)} - x^*\| \le \frac{\omega}{2}\,\|x^{(k)} - x^*\|^2\,.$$

If $0 < \|x^{(k)} - x^*\| \le \rho$, we have

$$\|x^{(k+1)} - x^*\| \le \underbrace{\frac{\omega}{2}\,\|x^{(k)} - x^*\|}_{\le\,\rho\omega/2<1}\,\|x^{(k)} - x^*\| < \|x^{(k)} - x^*\|\,,$$

whence $x^{(k)} \in K_\rho(x^*)\,,\ k \in \mathbb{N}$, and thus the convergence of the sequence $(x^{(k)})_{k\in\mathbb{N}}$ to x^* .

For the proof of uniqueness, let $x^{**} \in K_{2/\omega}(x^*)$ be another solution. Then, we have:

$$\|x^{**} - x^*\| = \|(F'(x^*))^{-1} [\underbrace{F(x^{**})}_{=0} - \underbrace{F(x^*)}_{=0} - F'(x^*)(x^{**} - x^*)]\| \le$$

$$\le \underbrace{\frac{\omega}{2} \|x^{**} - x^*\|}_{<1} \|x^{**} - x^*\| < \|x^{**} - x^*\| ,$$

whence $x^{**} = x^*$.

Remark 3.18 Monotonicity test

Idea: Use the residual $F(x^{(k)})$ to check for convergence (Requirement: Residual decreases monotonically). The monotonicity test

$$\|F(x^{(k+1)})\| \leq \bar{\Theta} \, \|F(x^{(k)})\|$$

with some $\bar{\Theta} < 1$ is not affine invariant.

Instead we use the affine invariant monotonicity test

$$(\bullet) \quad \|(F'(x^{(k)}))^{-1} \, F(x^{(k+1)})\| \leq \bar{\Theta} \, \|(F'(x^{(k)}))^{-1} \, F(x^{(k)})\| \, .$$

Since $-(F'(x^{(k)}))^{-1} \, F(x^{(k)})$ represents the Newton increment $\Delta x^{(k)} = x^{(k+1)} - x^{(k)}$, the implementation of the monotonicity test only requires the additional computation of $(F'(x^{(k)}))^{-1} \, F(x^{(k+1)})$. Denoting the solution $\bar{\Delta} x^{(k+1)}$ of

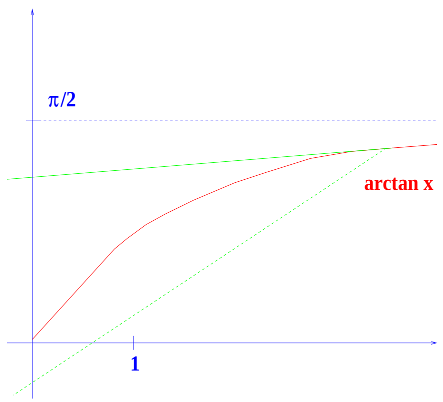
$$(F'(x^{(k)})) \, \bar{\Delta} x^{(k+1)} = - \, F(x^{(k+1)})$$

as the simplified Newton correction, the monotonicity test is as follows;

$$\|\bar{\Delta} x^{(k+1)}\| \leq \bar{\Theta} \, \|\Delta x^{(k)}\| \, .$$

In practice, the choice $\bar{\Theta} = 1/2$ has proven to be an efficient one.

Remark 3.19 Affine invariant damping strategy



In case the monotonicity test $(\bullet)$ fails, convergence can be often achieved by an appropriate damping of the Newton correction according to

$$(\diamond) \quad x^{(k+1)} \; = \; x^{(k+1)}(\lambda_k) \; := \; x^{(k)} \; + \; \lambda_k \, \Delta x^{(k)} \; .$$

Using the monotonicity test, the damping parameter λ_k with $0 < \lambda_k \leq 1$ can be computed by the following damping strategy:

For $\lambda_k \in \{1, \frac{1}{2}, \frac{1}{4}, ..., \lambda_{min}\}$ compute $\bar{\Delta} x^{(k+1)}(\lambda_k)$ as the solution of

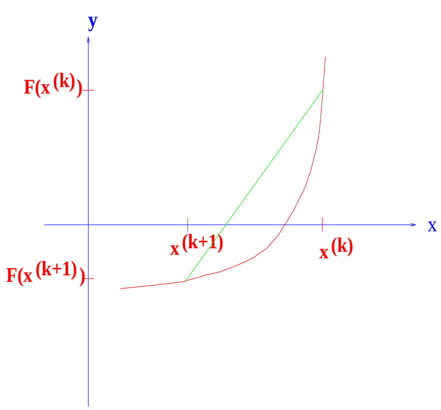
$$F'(x^{(k)}) \; \bar{\Delta} x^{(k+1)}(\lambda_k) \; = \; - \; F(x^{(k+1)}(\lambda_k))$$

and check whether

$$(\star) \quad \|\bar{\Delta} x^{(k+1)}(\lambda_k)\| \; \leq \; (1 \; - \; \frac{\lambda_k}{2}) \; \|\Delta x^{(k)}\| \; .$$

In case $(\star)$ is satisfied, $\lambda_{k+1} = \min(1, 2\lambda_k)$ is chosen in the next step. On the other hand, if $(\star)$ does not hold true, the test is repeated with $\tilde{\lambda}_k = \max(\lambda_{min}, \frac{\lambda_k}{2})$.

3.4 Quasi-Newton methods



Idea: Local affine approximation of $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$

$$S_{k+1}(x) \ := \ F(x^{(k+1)}) \ + \ B_{k+1}(x \ - \ x^{(k+1)}) \quad .$$

Specification of $B_{k+1} \in \mathbb{R}^{n \times n}$ such that

$$S_{k+1}(x^{(k)}) = F(x^{(k)}) \quad \Longrightarrow$$

$$(*) \quad B_{k+1} \underbrace{(x^{(k+1)} \ - \ x^{(k)})}_{=: p^{(k)}} \ = \ \underbrace{F(x^{(k+1)}) \ - \ F(x^{(k)})}_{=: y^{(k)}} \quad .$$

The condition $(*)$ is called the secant condition.

For $n \geq 2$, the matrix $B_{k+1} \in \mathbb{R}^{n \times n}$ is not uniquely determined. We have:

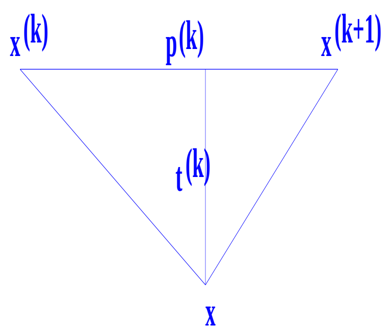
$$Q(p^{(k)}, y^{(k)}) \ := \ \{ \, B_{k+1} \in \mathbb{R}^{n \times n} \mid B_{k+1} p^{(k)} \ = \ y^{(k)} \, \} \quad , \quad \dim Q(p^{(k)}, y^{(k)}) \ = \ (n \ - \ 1) \, n \ .$$

Criterion for the selection: Least change in the affine model

$$S_{k+1}(x) - S_k(x) = (B_{k+1} - B_k)(x - x^{(k)}) .$$

Choose $B_{k+1} \in Q(p^{(k)}, y^{(k)})$ such that

$$(+) \quad \|B_{k+1} - B_k\|_F = \min_{B \in Q(p^{(k)}, y^{(k)})} \|B - B_k\|_F \quad , \quad \|B\|_F := \Big(\sum_{i,j=1}^{n} b_{ij}^2\Big)^{1/2} \quad \text{(Frobenius norm)} .$$



$$x - x^{(k)} = \alpha\, p^{(k)} + t^{(k)} \quad , \quad t^{(k)} \perp p^{(k)} \quad \Longrightarrow$$

$$(\diamond) \quad S_{k+1}(x) - S_k(x) = \underbrace{\alpha\,(B_{k+1} - B_k)\, p^{(k)}}_{=\alpha(y^{(k)} - B_k p^{(k)})} + (B_{k+1} - B_k)\, t^{(k)} \ .$$

Choose $B_{k+1} \in Q(p^{(k)}, y^{(k)})$ such that $(B_{k+1} - B_k)t^{(k)} = 0$.

$$\Longrightarrow \quad \text{Rang}\,(B_{k+1} - B_k) = 1 \quad \Longrightarrow \quad B_{k+1} - B_k = v^{(k)}(p^{(k)})^T$$

Using this in $(\diamond)$: $\alpha\, v^{(k)}\, (p^{(k)})^T\, p^{(k)} = \alpha\, (y^{(k)} - B_k p^{(k)}) \quad \Longrightarrow \quad v^{(k)} = (y^{(k)} - B_k p^{(k)})/(p^{(k)})^T p^k$.

Broyden's rank 1 update ("Good Broyden"):

$$(\star) \quad B_{k+1} = B_k + \big[\, F(x^{(k+1)}) - F(x^{(k)}) - B_k p^{(k)} \,\big]\, \frac{(p^{(k)})^T}{(p^{(k)})^T p^{(k)}} .$$

Update of the inverse by Sherman-Morrison-Woodbury formula:

$$(A + uv^T)^{-1} = A^{-1} - A^{-1} u v^T A^{-1} / (1 + v^T A^{-1} u) .$$

Setting $A := B_k$, $u := [F(x^{(k+1)}) - F(x^{(k)}) - Bp^{(k)}]$, $v := (p^{(k)})^T/(p^{(k)})^T p^{(k)}$, we obtain:

$$B_{k+1}^{-1} = B_k^{-1} + \frac{[p^{(k)} - B_k^{-1}(F(x^{(k+1)}) - F(x^{(k)}))] (p^{(k)})^T B_k^{-1}}{(p^{(k)})^T B_k^{-1} [F(x^{(k+1)}) - f(x^{(k)})]} .$$

Alternative selection: Least change in the solution of the affine model

$$\|B_{k+1}^{-1} - B_k^{-1}\|_F = \min_{B \in Q(p^{(k)}, y^{(k)})} \|B^{-1} - B_k^{-1}\|_F .$$

$\Longrightarrow$ Alternative rank 1 update (”Bad Broyden”):

$$(\bullet) \quad B_{k+1}^{-1} = B_k^{-1} + \frac{[p^{(k)} - B_k^{-1}(F(x^{(k+1)}) - F(x^{(k)}))] (F(x^{(k+1)}) - F(x^{(k)}))^T}{(F(x^{(k+1)}) - F(x^{(k)}))^T (F(x^{(k+1)}) - F(x^{(k)}))} .$$

3.5 Nonlinear least squares problems

Example 3.20 Kinetics of coupled chemical reactions

A chemical substance S_1 is injected into a reaction chamber. Due to high temperature or catalytics, reactions with substances S_2 , S_3 occur:

$$S_1 \rightarrow S_2 \quad , \quad S_2 \rightarrow S_3$$

We assume that the concentrations $[S_i]$, $1 \leq i \leq 3$, behave according to

$$\frac{d}{dt}[S_1] \sim -[S_1] \quad , \quad \frac{d}{dt}[S_2] \sim [S_1], [S_2] \quad , \quad \frac{d}{dt}[S_3] \sim [S_2] .$$

The constants $k_2 > k_1 \geq 0$ are the reaction velocities of both reactions.

We thus obtain the following system of differential equations

$$\frac{d}{dt}[S_1] = -k_1[S_1] \quad , \quad \frac{d}{dt}[S_2] = k_1[S_1] - k_2[S_2] \quad , \quad \frac{d}{dt}[S_3] = k_2[S_2]$$

with the initial conditions: $[S_1](0) = S_1^0$, $[S_2](0) = [S_3](0) = 0$.

By elementary integration we obtain:

$$[S_1](t) \;=\; S_1^0 \exp(-\,k_1\,t)\ ,$$
$$[S_2](t) \;=\; \frac{k_1\,S_1^0}{k_2\,-\,k_1}\,[\,\exp(-\,k_1\,t)\;-\;\exp(-\,k_2\,t)\,]\ ,$$
$$[S_3](t) \;=\; S_1^0\,[\,1\;-\;\frac{k_2}{k_2\,-\,k_1}\exp(-\,k_1\,t)\;-\;\frac{k_1\,S_1^0}{k_2\,-\,k1}\exp(-\,k_2\,t)\,]\ .$$

Problem: Determination of the initial concentrations S_1^0 and of the reaction velocities $k_1\ ,\ k_2$ by measurements $b_i\ ,\ 1 \leq i \leq m\ ,\ m \geq 3$, of the concentrations $[S_2](t_i)$.

Compute $S_2^0\ ,\ k_1\ ,\ k_2$ such that

$$\sum_{i=1}^{m} (\,b_i\;-\;[S_2](t_i)\,)^2$$

is minimized. .

Definition 3.21 Nonlinear least squares problem

Assume that $\varphi = \varphi(t; x_1, ..., x_n)$ is a parameter dependent function depending on t and $b_i \in \mathbb{R}$, $1 \leq i \leq m$. Further, let $F : D \subset \mathbb{R}^n \to \mathbb{R}^m$ be the (nonlinear) function given by

$$F_i(x) \ := \ b_i \ - \ \varphi(t_i; x_1, ..., x_n) \quad , \quad 1 \leq i \leq m$$

The problem to determine $x^* = (x_1^*, ..., x_n^*)^T \in D$ such that

$$(*) \quad g(x^*) \ := \ \|F(x^*)\|^2 \ = \ \min_{x \in D} \|F(x)\|^2 \ ,$$

is said to be a nonlinear least squares problem.

Remark 3.22 Local inner minima

The sufficient conditions for a local inner minimum $x^* \in D$ von g are given by:

$$(\diamond) \quad g'(x^*) \ = \ 0 \quad , \quad g''(x^*) \text{ is positive definite} \ ,$$

where g'' is the Hessian (matrix of second partial derivatives) of g .

We obtain:

$$g(x) = F(x)^T F(x) ,$$
$$g'(x) = F'(x)^T F(x) + (F(x)^T F'(x))^T = 2 F'(x)^T F(x) .$$

Hence, $g'(x) = 0$ holds true if and only if

$$(+) \quad G(x) := F'(x)^T F(x) = 0 .$$

3.5.1 Gauss Newton method

Remark 3.23 Newton iteration

Newton's method applied to $(+)$ results in:

$$G'(x^{(k)}) \, \Delta x^{(k+1)} = - G(x^{(k)}) \quad , \quad k \in \mathbb{N}_0 ,$$
$$G'(x) = F'(x)^T F'(x) + F''(x)^T F(x) .$$

In view of $(\diamond)$, $G'(x^*)$ is positive definite and hence, it is invertible in some neighborhood $U_\delta(x^*)$. $G'(x)$ is positive definite if and only if $\mathrm{rank}\, F'(x) = n$ (maximum rank).

Definition 3.24 Compatibility
The nonlinear least squares problem $(*)$ is said to be compatible, if $F(x^*) = 0$ and almost compatible, if $\|F(x^*)\| \ll 1$.

Remark 3.25 Consequences of compatibility

If the nonlinear least squares problem is compatible, there holds

$$G'(x^*) \;=\; F'(x^*)^T \, F'(x^*) \; .$$

In case of a compatible or almost compatible least squares problem, we may neglect the term $F''(x)^T \, F(x)$ in $G'(x)$. This leads to the iteration:

$$(\star) \quad F'(x^{(k)})^T \, F'(x^{(k)}) \, \Delta x^{(k+1)} \;=\; -\, F'(x^{(k)})^T \, F(x^{(k)}) \quad , \quad k \in \mathbb{N}_0 \; .$$

Definition 3.26 Gauss Newton method
The iteration $(\star)$ is called the Gauss Newton method for the nonlinear least squares problem $(*)$.

Remark 3.27 Normal equations
Setting $A := F'(x^{(k)})$ and $b := -F(x^{(k)})$, $(\star)$ corresponds to the normal equations $A^T A \; \Delta x^{(k+1)} = A^T b$ of the linear least squares problem $\min \|b - A \; \Delta x\|$.

Theorem 3.28 Properties of the Gauss Newton method

(i) Assume that the Jacobi matrix $F'(x^*)$ has maximum rank $\text{rank } F'(x^*) = n$. Then, the nonlinear least squares problem $(*)$ admits a unique solution $x^* \in D$.

(ii) In case of $\text{rank } F'(x^*) < n$, there exists a manifold of solutions of the dimension $n - \text{rank } F'(x^*)$ (rank defect).

(iii) If the nonlinear least squares problem is compatible and if $F'(x^*)$ has maximum rank, the Gauss Newton method converges locally quadratically.

(iv) If the nonlinear least squares problem is almost compatible and if $F'(x^*)$ has maximum rank, we have locally linear convergence.

Remark 3.29 Damping strategy

As in case of nonlinear systems of equations, the area of convergence can be extended by an appropriate damping strategy:

Denoting by $\Delta x^{(k+1)}$ the Gauss Newton correction computed according to $(\star)$, a new iterate

$$x^{(k+1)} = x^{(k)} + \lambda_k \Delta x^{(k+1)}$$

is computed where $0 < \lambda_k \leq 1$ is a suitably chosen damping parameter.

3.5.2 Trust-region method (Levenberg-Marquardt algorithm)

Idea: Given the iterate $x^{(k)}$, approximate the nonlinear function F by the affine function

$$F(x^{(k)}) + F'(x^{(k)})(x - x^{(k)})$$

and compute $x^{(k+1)} = x^{(k)} + \Delta x^{(k+1)}$ as the solution of the linear least squares problem

$$\min_{\Delta x} \|F(x^{(k)}) + F'(x^{(k)})\,\Delta x\|$$

in a suitable neighborhood $T(x^{(k)}) = \{\Delta x \in \mathbb{R}^n \,|\, \|\Delta x\| \le \Delta_k\}$ (trust region) of $x^{(k)}$.

This leads to the constrained linear least squares problem:

$$\min_{\Delta x \in T(x^{(k)})} \|F(x^{(k)}) + F'(x^{(k)})\,\Delta x\|$$

The coupling of the constraint by a Lagrange multiplier $\lambda_k \ge 0$ gives rise to the saddle point problem:

$$(\bullet) \quad \min_{\Delta x \in \mathbb{R}^n} \max_{\lambda_k \ge 0} \left[\|F(x^{(k)}) + F'(x^{(k)})\,\Delta x\|^2 + \lambda_k \left(\|\Delta x\|^2 - \Delta_k^2\right) \right]$$

The optimality conditions for the saddle point problem

$$\min_{\Delta x \in \mathbb{R}^n} \max_{\lambda_k \geq 0} \left[\|F(x^{(k)}) + F'(x^{(k)}) \Delta x\|^2 + \lambda_k (\|\Delta x\|^2 - \Delta_k^2) \right]$$

are given by the linear complementarity problem

$$(1) \qquad [(F'(x^{(k)}))^T F'(x^{(k)}) + \lambda_k I] \Delta x = - F'(x^{(k)})^T F(x^{(k)}) ,$$

$$(2) \qquad \left. \begin{array}{c} \lambda_k \geq 0 \\ \Delta_k \geq \|\Delta x\| \end{array} \right\} \lambda_k (\|\Delta x\| - \Delta_k) = 0 .$$

Definition 3.30 Levenberg-Marquardt algorithm

The method $(\bullet)$ resp. (1), (2) is called the Levenberg-Marquardt algorithm.

Remark 3.31 Pros and cons of the Levenberg-Marquardt algorithm

Pro: For $\lambda_k > 0$, the coefficient matrix in (1) is positive definite, even if $F'(x^{(k)})$ is singular.

Con: There is no information about a rank defect of $F'(x^{(k)})$. Hence, the uniqueness of the nonlinear least squares problem remains open. The algorithm is not affine invariant, and the control of the Lagrange multiplier λ_k is a difficult task.