Numerical Methods for Large-Scale Nonlinear Systems

Handouts by Ronald H.W. Hoppe

following the monograph

P. Deuflhard

Newton Methods for Nonlinear Problems

Springer, Berlin-Heidelberg-New York, 2004

1. Classical Newton Convergence Theorems

1.1 Classical Newton-Kantorovich Theorem

Theorem 1.1 Classical Newton-Kantorovich Theorem

Let X and Y be Banach spaces, $D \subset X$ a convex subset, and suppose that that $F : D \subset X \to Y$ is continuously Fréchet differentiable on D with an invertible Fréchet derivative $F'(x^0)$ for some initial guess $x^0 \in D$. Assume further that the following conditions hold true:

$$\|F'(x^0)^{-1}F(x^0\| \;\leq\; \alpha \;, \tag{1.1}$$

$$\|F'(y) - F'(x)\| \;\leq\; \gamma \;\|y - x\| \quad , \quad x, y \in D \;, \tag{1.2}$$

$$h_0 \;:=\; \alpha\;\gamma\;\|F'(x^0)^{-1}\| \;<\; \frac{1}{2} \;, \tag{1.3}$$

$$\overline{B}(x^0, \rho_0) \subset D \;,\; \rho_0 \;:=\; \frac{1 - \sqrt{1 - 2h_0}}{\gamma\;\|F'(x^0)^{-1}\|} \;. \tag{1.4}$$

Then, for the sequence $\{x^k\}_{\mathbb{N}_0}$ of Newton iterates

$$\begin{aligned} F'(x^k)\;\Delta x^k \;&=\; -\;F(x^k) \;, \\ x^{k+1} \;&=\; x^k \;+\; \Delta x^k \end{aligned}$$

there holds

(i) $F'(x)$ is invertible for all Newton iterates $x = x^k, k \in \mathbb{N}_0$,

(ii) The sequence $\{x^k\}_{\mathbb{N}}$ of Newton iterates is well defined with $x^k \in \overline{B}(x^0, \rho_0)$, $k \in \mathbb{N}_0$, and $x^k \to x^* \in \overline{B}(x^0, \rho_0), k \in \mathbb{N}_0$ $(k \to \infty)$, where $F(x^*) = 0$,

(iii) The convergence $x^k \to x^*$ $(k \to \infty)$ is quadratic,

(iv) The solution x^* of $F(x) = 0$ is unique in

$$\overline{B}(x^0, \rho_0) \cup (D \cap B(x^0, \overline{\rho}_0)) \quad , \quad \overline{\rho}_0 \;:=\; \frac{1 + \sqrt{1 - 2h_0}}{\gamma\;\|F'(x^0)^{-1}\|} \;.$$

Proof. We have

$$\|F'(x^k)\;-F'(x^0)\| \;\leq\; \gamma\;\|x^k - x^0\| \;\leq\; t_k$$

for some upper bound $t_k, k \in \mathbb{N}$.
If we can prove $x^k \in B(x^0, \rho_0)$ and $\tilde{t}_k := \|F'(x^0)^{-1}\|t_k < 1, k \in \mathbb{N}$, then by the

Banach perturbation lemma $F'(x^k)$ is invertible with

$$\begin{aligned}\|F'(x^k)^{-1}\| &\leq \frac{\|F'(x^0)^{-1}\|}{1-\|F'(x^0)^{-1}\|\|F'(x^k)-F'(x^0)\|} \leq \\ &\leq \frac{\|F'(x^0)^{-1}\|}{1-\gamma\|F'(x^0)^{-1}\|\|x^k-x^0\|} \leq \frac{\|F'(x^0)^{-1}\|}{1-\tilde{t}_k} =: \beta_k \ .\end{aligned} \tag{1.5}$$

We prove $x^k \in B(x^0,\rho_0)$ and $\tilde{t}_k < 1, k \in \mathbb{N}$, by induction on k:
For $k=1$ we have

$$\|x^1-x^0\| = \|F'(x^0)^{-1}F(x^0)\| \leq \alpha = \frac{h_0}{\gamma\|F'(x^0)^{-1}\|} < \rho_0 \ ,$$

since $h_0 < 1-\sqrt{1-2h_0}$, and

$$\begin{aligned}\tilde{t}_1 &:= \|F'(x^0)^{-1}\|\ t_1 = \gamma\ \|F'(x^0)^{-1}\|\ \|x^1-x^0\| = \\ &= \gamma\ \|F'(x^0)^{-1}\|\ \|F'(x^0)^{-1}F(x^0)\| \leq \underbrace{\alpha\ \gamma\ \|F'(x^0)^{-1}\|}_{=\ h_0} < \frac{1}{2} \ .\end{aligned}$$

Assuming the assertion to be true for some $k \in \mathbb{N}$, for $k+1$, using (1.2) we obtain

$$\begin{aligned}&\|x^{k+1}-x^k\| = \|F'(x^k)^{-1}F(x^k)\| = \\ &= \|F'(x^k)^{-1}\ (F(x^k)-F(x^{k-1})-F'(x^{k-1})\Delta x^{k-1})\| = \\ &= \|F'(x^k)^{-1}\int_0^1 \Big(F'(x^{k-1}+s\Delta x^{k-1})-F'(x^{k-1})\Big)\Delta x^{k-1}\ ds\| \leq \\ &\leq \|F'(x^k)^{-1}\|\int_0^1 \underbrace{\|F'(x^{k-1}+s\Delta x^{k-1})-F'(x^{k-1})\|}_{\leq\ s\ \gamma\ \|\Delta x^{k-1}\|}\ \|\Delta x^{k-1}\|\ ds \leq \\ &\leq \frac{1}{2}\ \beta_k\ \gamma\ \|x^k-x^{k-1}\|^2 \ .\end{aligned} \tag{1.6}$$

Setting

$$h_k := \gamma\ \|x^{k+1}-x^k\| \ ,$$

we thus get the recursion

$$h_k = \frac{1}{2}\ \beta_k\ h_{k-1}^2 \quad , \quad k \in \mathbb{N} \ . \tag{1.7}$$

In view of the relationship

$$\underbrace{\gamma\ \|x^{k+1}-x^0\|}_{\leq\ t_{k+1}} \leq \underbrace{\gamma\ \|x^{k+1}-x^k\|}_{=\ h_k} + \underbrace{\gamma\ \|x^k-x^0\|}_{\leq\ t_k} \ ,$$

we consider the recursion

$$t_{k+1} \;=\; t_k \;+\; h_k \;. \tag{1.8}$$

Observing (1.5) and (1.7), we find

$$t_{k+1} - t_k \;=\; \frac{1}{2}\,\frac{\|F'(x^0)^{-1}\|}{1-\tilde{t}_k}\,(t_k - t_{k-1})^2 \;.$$

Hence, multiplying both sides with $\|F'(x^0)^{-1}\|$, we end up with the following three-term recursion for $\tilde{t}_k$:

$$\tilde{t}_{k+1} - \tilde{t}_k \;=\; \frac{1}{2}\,\frac{1}{1-\tilde{t}_k}\,(\tilde{t}_k - \tilde{t}_{k-1})^2 \quad , \quad \tilde{t}_0 = 0 \;,\; \tilde{t}_1 = h_0 \;. \tag{1.9}$$

The famous **Ortega-trick** allows to reduce (1.9) to a two-term recursion which can be interpreted as a Newton method in $\mathbb{R}^1$:
Multiplying both sides in (1.9) by $1-\tilde{t}_k$ results in

$$(\tilde{t}_{k+1} - \tilde{t}_k)\,(1-\tilde{t}_k) \;=\; \frac{1}{2}\,(\tilde{t}_k - \tilde{t}_{k-1})^2 \;,$$

from which we deduce

$$\underbrace{\tilde{t}_{k+1} - \tilde{t}_{k+1}\tilde{t}_k + \frac{1}{2}\tilde{t}_k^2}_{=\,\psi(\tilde{t}_{k+1},\tilde{t}_k)} \;=\; \underbrace{\tilde{t}_k - \tilde{t}_k\tilde{t}_{k-1} + \frac{1}{2}\tilde{t}_{k-1}^2}_{=\,\psi(\tilde{t}_k,\tilde{t}_{k-1})} \;.$$

It follows that

$$\psi(\tilde{t}_{k+1},\tilde{t}_k) \;=\; \psi(\tilde{t}_1,\tilde{t}_0) \;=\; h_0 \;,$$

from which we deduce

$$\tilde{t}_{k+1} - \tilde{t}_k \;=\; \frac{h_0 - \tilde{t}_k + \frac{1}{2}\tilde{t}_k^2}{1-\tilde{t}_k} \;=\; -\,\frac{\varphi(\tilde{t}_k)}{\varphi'(\tilde{t}_k)} \;,$$

where $\varphi : \mathbb{R} \to \mathbb{R}$ is given by

$$\varphi(\tilde{t}) \;:=\; h_0 - \tilde{t} + \frac{1}{2}\tilde{t}^2 \;.$$

Obviously, φ has the zeroes

$$\tilde{t}_1^* \;:=\; 1 - \sqrt{1-2h_0} \quad , \quad \tilde{t}_2^* \;:=\; 1 + \sqrt{1-2h_0} \;.$$

Since φ is convex, the Newton method converges for $\tilde{t}_1$ to $\tilde{t}_1^*$.
It follows from the definition of $\tilde{t}_k$ that $x^k \in B(x^0, \rho_0)$. Moreover, as a consequence of (1.6) we readily find that $\{x^k\}_{\mathbb{N}}$ is a Cauchy sequence in $\overline{B}(x^0, \rho_0)$. Hence, there exists $x^* \in \overline{B}(x^0, \rho_0)$ such that $x^k \to x^*$ $(k \to \infty)$ and

$$\Delta x^k \;=\; -F'(x^k)^{-1}F(x^k) \to -\,F'(x^*)^{-1}F(x^*) \;=\; 0\;,$$

whence $F(x^*) = 0$.
The quadratic convergence can be deduced from (1.6) as well. Finally, the uniqueness of x^* in $\overline{B}(x^0, \rho_0) \cup (D \cap B(x^0, \overline{\rho}_0))$ follows readily from the properties of the function φ.

•

1.2 Classical Newton-Mysovskikh Theorem

Theorem 1.2 Classical Newton-Mysovskikh Theorem

Let X and Y be Banach spaces, $D \subset X$ a convex subset, and suppose that that $F : D \subset X \to Y$ is continuously Fréchet differentiable on D with invertible Fréchet derivatives $F'(x), x \in D$, and let $x^0 \in D$ be some initial guess. Assume further that the following conditions hold true:

$$\|F'(x^0)^{-1}F(x^0\| \;\le\; \alpha\;, \tag{1.10}$$

$$\|F'(x)^{-1}\| \;\le\; \beta \quad,\quad x \in D\;, \tag{1.11}$$

$$\|F'(y) - F'(x)\| \;\le\; \omega\,\|y - x\| \quad,\quad x, y \in D\;, \tag{1.12}$$

$$h_0 \;:=\; \frac{1}{2}\,\beta\,\omega\,\|F'(x^0)^{-1}F(x^0\| \;\le\; \frac{1}{2}\,\alpha\,\beta\,\omega \;<\; 1\;, \tag{1.13}$$

$$\overline{B}(x^0, \rho) \subset D\;,\; \rho \;:=\; \alpha \sum_{j=0}^{\infty} h_0^{2^j-1} \;\le\; \frac{\alpha}{1-h_0}\;. \tag{1.14}$$

Then, for the sequence $\{x^k\}_{\mathbb{N}_0}$ of Newton iterates

$$\begin{aligned} F'(x^k)\,\Delta x^k \;&=\; -\,F(x^k)\;,\\ x^{k+1} \;&=\; x^k \;+\; \Delta x^k \end{aligned}$$

there holds

(i) $x^k \in \overline{B}(x^0, \rho), k \in \mathbb{N}_0$, and there exists $x^* \in \overline{B}(x^0, \rho)$ such that $F(x^*) = 0$ and $x^k \to x^*$ $(k \to \infty)$,

(ii) $$\|x^{k+1} - x^k\| \;\le\; \tfrac{1}{2}\,\beta\,\omega\,\|x^k - x^{k-1}\| \quad,\quad k \in \mathbb{N},$$

(iii) $$\|x^k - x^*\| \;\le\; \varepsilon_k\,\|x^k - x^{k-1}\|^2 \quad,\text{where}$$

$$\varepsilon_k \;:=\; \frac{1}{2}\,\beta\,\omega\,(1 + \sum_{j=1}^{\infty}(h_0^{2^k})^{2^j}) \;\le\; \frac{1}{2}\,\frac{\beta\,\omega}{1-h_0^{2^k}}\;.$$

Proof. Observing

$$F'(x^{k-1})\ \Delta x^{k-1}\ +\ F(x^{k-1})\ =\ 0\ ,$$

we obtain

$$\begin{aligned}
&\|\Delta x^k\| \ =\ \|F'(x^k)^{-1}F(x^k)\| \ = \\
&=\ \|F'(x^k)^{-1}\ \Big(F(x^k)-F(x^{k-1})-F'(x^{k-1})\Delta x^{k-1}\Big)\|\ \leq \\
&\leq\ \|F'(x^k)^{-1}\|\ \|\int_0^1 \Big(F'(x^{k-1}+s\Delta x^{k-1})-F'(x^{k-1})\Big)\Delta x^{k-1}\ ds\|\ \leq \\
&\leq\ \frac{1}{2}\ \beta\ \omega\ \|\Delta x^{k-1}\|^2\ ,
\end{aligned}$$

which gives the assertion **(ii)**.
We now prove that $\{x^k\}_{\mathbb{N}_0}$ is a Cauchy sequence in $\overline{B}(x^0,\rho)$. By induction on k we show

$$\|x^{k+1}-x^k\|\ \leq\ \frac{2}{\beta\omega}\ h_0^{2^k}\quad ,\quad k\in\mathbb{N}_0\ . \tag{1.15}$$

For $k=0$, we have in view of (1.10)

$$\|\Delta x^0\|\ \leq\ \alpha\ =\ \frac{2}{\beta\omega}\ h_0\ .$$

Assuming (1.15) to be true for some $k\in\mathbb{N}$, we get

$$\begin{aligned}
&\|x^{k+2}-x^{k+1}\|\ =\ \|\Delta x^{k+1}\|\ \leq\ \frac{1}{2}\ \beta\ \omega\ \|\Delta x^k\|^2\ \leq \\
&\leq\ \frac{1}{2}\ \beta\ \omega\ \Big(\frac{2}{\beta\omega}\ h_0^{2^k}\Big)^2\ =\ \frac{2}{\beta\omega}\ h_0^{2^{k+1}}\ .
\end{aligned}$$

It follows readily from (1.15) that $x^{k+1}\in\overline{B}(x^0,\rho)$:

$$\begin{aligned}
&\|x^{k+1}-x^0\|\ \leq\ \|x^{k+1}-x^k\|\ +\ \ldots\ +\ \|x^1-x^0\|\ \leq \\
&\leq\ \frac{2}{\beta\omega}\ \Big(h_0^{2^k}\ +\ \ldots\ +\ h_0\Big)\ =\ \underbrace{\frac{2}{\beta\omega}\ h_0}_{=\ \alpha}\ \sum_{j=0}^{\infty} h_0^{2^j-1}\ =\ \rho\ .
\end{aligned}$$

Similarly, it can be shown that

$$\|x^{m+k}-x^m\|\ \to\ 0\quad (,\to\infty)\ .$$

Since $\{x^k\}_{\mathbb{N}_0}$ is a Cauchy sequence in $\overline{B}(x^0,\rho)$, there exists $x^* \in \overline{B}(x^0,\rho)$ such that $x^k \to x^*$ $(k \to \infty)$. Hence,

$$\Delta x^k \;=\; -\,F'(x^k)^{-1}F(x^k) \;\to\; -\,F'(x^*)F(x^*) \;=\; 0\;,$$

and thus $F(x^*)=0$ which proves **(i)**.
The assertion **(iii)** is shown as follows: Setting

$$h_k \;:=\; \frac{1}{2}\,\beta\,\omega\,\|\Delta x^k\|\;,$$

we obtain

$$\begin{aligned}
&\|x^k - x^*\| \;=\; \lim_{k<m\to\infty}\|x^k - x^m\| \;\le\\
&\le\; \lim_{k<m\to\infty}\Big[\,\|x^m - x^{m-1}\| \;+\; \ldots \;+\; \|x^{k+1}-x^k\|\,\Big] \;\le\\
&\le\; \frac{2}{\beta\omega}\;\lim_{k<m\to\infty}\Big[\,h_{m-1} \;+\; \ldots \;+\; h_k\,\Big] \;=\\
&=\; \frac{2h_k}{\beta\omega}\;\lim_{k<m\to\infty}\Big[\,1 \;+\; \frac{h_{k+1}}{h_k} \;+\; \ldots \;+\; \frac{h_{m-1}}{h_k}\,\Big]\;.
\end{aligned}$$

On the other hand, taking **(ii)** into account

$$h_k \;\le\; \Big(\frac{\beta\omega}{2}\Big)^2\,\|\Delta x^{k-1}\|^2 \;=\; h_{k-1}^2\;,$$

whence

$$h_{k+\ell} \;\le\; h_k^{2^\ell} \quad,\quad k\in\mathbb{N}_0\;.$$

We conclude

$$\begin{aligned}
&\|x^k - x^*\| \;\le\; \frac{1}{2}\,\beta\,\omega\,\|\Delta x^{k-1}\|^2\,\Big[\,1 \;+\; h_k^2 \;+\; \ldots\Big] \;\le\\
&\le\; \frac{1}{2}\,\beta\,\omega\,\Big(1 \;+\; \sum_{j=1}^{\infty} h_k^{2^j}\Big)\,\|\Delta x^{k-1}\|^2\;,
\end{aligned}$$

which proves **(iii)**.

•

2. Affine Invariant/Conjugate Newton Convergence Theorems

2.1 Affine Covariant Newton Convergence Theorems

Theorem 2.1 Affine Covariant Newton-Kantorovich Theorem

Let $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable on D with an invertible Jacobian $F'(x^0)$ for some initial guess $x^0 \in D$. Assume further that the following conditions hold true:

$$\|F'(x^0)^{-1} F(x^0\| \;\leq\; \alpha \;, \tag{1.16}$$

$$\|F'(x^0)^{-1}\; (F'(y) - F'(x))\| \;\leq\; \gamma\; \|y - x\| \quad , \quad x, y \in D \;, \tag{1.17}$$

$$h_0 \;:=\; \alpha\; \gamma \;<\; \frac{1}{2} \;, \tag{1.18}$$

$$\overline{B}(x^0, \rho_0) \subset D \;,\; \rho_0 \;:=\; \frac{1 - \sqrt{1 - 2h_0}}{\gamma} \;. \tag{1.19}$$

Then, for the sequence $\{x^k\}_{\mathbb{N}_0}$ of Newton iterates

$$\begin{aligned} F'(x^k)\; \Delta x^k \;&=\; -\; F(x^k) \;, \\ x^{k+1} \;&=\; x^k \;+\; \Delta x^k \end{aligned}$$

there holds

(i) $F'(x)$ is invertible for all Newton iterates $x = x^k, k \in \mathbb{N}_0$,

(ii) The sequence $\{x^k\}_{\mathbb{N}}$ of Newton iterates is well defined with $x^k \in \overline{B}(x^0, \rho_0)$, $k \in \mathbb{N}_0$, and $x^k \to x^* \in \overline{B}(x^0, \rho_0), k \in \mathbb{N}_0$ $(k \to \infty)$, where $F(x^*) = 0$,

(iii) The convergence $x^k \to x^*$ $(k \to \infty)$ is quadratic,

(iv) The solution x^* of $F(x) = 0$ is unique in

$$\overline{B}(x^0, \rho_0) \cup (D \cap B(x^0, \overline{\rho}_0)) \quad , \quad \overline{\rho}_0 \;:=\; \frac{1 + \sqrt{1 - 2h_0}}{\gamma} \;.$$

Proof. First homework assignment.

Theorem 2.2 Affine Covariant Newton-Mysovskikh Theorem

Let $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$, $d \subset \mathbb{R}^n$ convex, be continuously differentiable on D with invertible Jacobians $F'(x), x \in D$, and let $x^0 \in D$ be some initial guess. Assume further that the following conditions hold true:

$$\|F'(x^0)^{-1}F(x^0)\| \;\leq\; \alpha \;, \tag{1.20}$$

$$\|F'(z)^{-1}\Big(F'(y) - F'(x)\Big)(y-x)\| \;\leq\; \omega\, \|y-x\|^2 \quad , \quad x,y,z \in D \tag{1.21}$$

$$h_0 \;:=\; \omega\, \|\Delta x^0\| \;\leq\; \alpha\,\omega \;<\; 2 \;, \tag{1.22}$$

$$\overline{B}(x^0,\rho) \subset D \quad , \quad \rho := \frac{\|\Delta x^0\|}{1-\frac{h_0}{2}} \;. \tag{1.23}$$

Then, for the sequence $\{x^k\}_{\mathbb{N}_0}$ of Newton iterates

$$\begin{aligned} F'(x^k)\;\Delta x^k \;&=\; -\;F(x^k) \;, \\ x^{k+1} \;&=\; x^k \;+\; \Delta x^k \end{aligned}$$

there holds $x^k \in B(x^0,\rho), k \in \mathbb{N}_0$, and there exists $x^* \in \overline{B}(x^0,\rho)$ such that $F(x^*) = 0$ and $x^k \to x^*$ $(k \to \infty)$ with

$$\|x^{k+1} - x^k\| \;\leq\; \frac{1}{2}\,\omega\,\|x^k - x^{k-1}\|^2 \;,$$

$$\|x^k - x^*\| \;\leq\; \frac{\|x^k - x^{k-1}\|}{1 - \frac{1}{2}\,\omega\,\|x^k - x^{k-1}\|} \;.$$

Proof. slight modification of the Classical Newton-Mysovskikh Theorem. •

2.2 Affine Contravariant Newton Convergence Theorem

Theorem 2.3 Affine Contravariant Newton-Mysovskikh Theorem

Let $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$, $d \subset \mathbb{R}^n$ convex, be continuously differentiable on D with invertible Jacobians $F'(x), x \in D$, and let $x^0 \in D$ be some initial guess. Assume further that the following conditions hold true:

$$\|\Big(F'(y) - F'(x)\Big)(y-x)\| \;\leq\; \omega\, \|F'(x)(y-x)\|^2 \quad , \quad x,y \in D \;, \tag{1.24}$$

$$\overline{\mathcal{L}}_\omega \subset D \quad , \quad \mathcal{L}_\omega \;:=\; \{x \in D \mid \|F(x)\| < \frac{2}{\omega} \;, \tag{1.25}$$

$$h_0 \;:=\; \omega\, \|F(x^0\| \;<\; 2 \;. \tag{1.26}$$

Then, the sequence $\{x^k\}_{\mathbb{N}_0}$ of Newton iterates stays in $\mathcal{L}_\omega$, and there exists an $x^* \in \mathcal{L}_\omega$ such that $x^k \to x^*$ for some subsequence $N' \subset \mathbb{N}$ and $F(x^*) = 0$. Moreover, for the residuals $F(x^k)$ there holds

$$\|F(x^k)\| \;\leq\; \frac{1}{2}\,\omega\,\|F(x^k)\|^2\;.$$

Proof. We first prove $x^k \in \mathcal{L}_\omega$ by induction on k:

(i) $k = 0$: in view of (1.26)

$$\|F(x^0)\| \;<\; \frac{2}{\omega} \quad\Longrightarrow\quad x^0 \in \mathcal{L}_\omega\;.$$

(ii) Assume that the assertion holds true for some $k \in \mathbb{N}$.

(iii) For any $\lambda \in [0,1]$ such that $x^k + t\Delta x^k \in \mathcal{L}_\omega, t \in [0,\lambda]$, we have

$$\|F(x^k + \lambda\Delta x^k)\| \;=\; \|F(x^k) \;+\; \int_0^\lambda F'(x^k + t\Delta x^k)\Delta x^k\;dt\|\;.$$

Since $F(x^k) = -F'(x^k)\Delta x^k$,

$$F(x^k) \;=\; -\;\lambda\;F'(x^k)\Delta x^k \;+\; (1-\lambda)\;F(x^k)\;,$$

and hence,

$$\|F(x^k + \lambda\Delta x^k)\| \;= \tag{1.27}$$

$$=\; \|\int_0^\lambda \Big[\Big(F'(x^k + t\Delta x^k) - F'(x^k)\Big)\;\Delta x^k \;+\; (1-\lambda)\;F(x^k)\Big]\;dt\| \;\leq$$

$$\leq\; \int_0^\lambda \underbrace{\|\Big(F'(x^k + t\Delta x^k) - F'(x^k)\Big)\;\Delta x^k\|}_{\leq\;\omega t\|F'(x^k)\Delta x^k)\|^2}\;dt \;+\; (1-\lambda)\;\|F(x^k)\| \;\leq$$

$$\leq\; \omega\int_0^\lambda \|\underbrace{F'(x^k)\Delta x^k}_{=\;-F(x^k)}\|^2\;t\;dt \;+\; (1-\lambda)\;\|F(x^k)\| \;=$$

$$= \ (1 - \lambda + \frac{1}{2}\omega\lambda^2\|F(x^k)\|) \ \|F(x^k)\| \ .$$

We assume

$$x^{k+1} \ = \ x^k \ + \ \Delta x^k \notin \mathcal{L}_\omega \ .$$

Then there exists

$$\overline{\lambda} \ := \ \min\{\lambda \in (0,1] \mid x^k \ + \ \lambda\Delta x^k \in \partial\mathcal{L}_\omega\} \ .$$

It follows from (1.27)

$$\|F(x^k + \overline{\lambda}\Delta x^k)\| \ \leq$$

$$\leq \ (1 - \overline{\lambda} + \frac{1}{2}\omega\overline{\lambda}^2 \ \underbrace{\|F(x^k)\|}_{< \frac{2}{\omega}}) \ \|F(x^k)\| \ <$$

$$< \ \underbrace{(1 - \overline{\lambda} + \overline{\lambda}^2)}_{< 1} \ \underbrace{\|F(x^k)\|}_{< \frac{2}{\omega}} \ < \ \frac{2}{\omega} \ ,$$

and hence, $x^k + \overline{\lambda}\Delta x^k \in \mathcal{L}_\omega$ which is a contradiction.
For $\lambda = 1$, (1.27) gives the asserted residual estimate.
For the proof of the rest of the assertion, we define the residual oriented Kantorovich quantities

$$h_k \ := \ \omega \ \|F(x^k)\| \ .$$

then, (1.26) implies

$$\omega \ \|F(x^{k+1})\| \ \leq \ \frac{1}{2} \ \omega^2 \ \|F(x^k)\| \ ,$$

i.e.,

$$h_{k+1} \ \leq \ \frac{1}{2} \ h_k^2 \ = \ \frac{1}{2} h_k \ h_k \ .$$

Since $h_0 < 2$, for $k = 0$ we obtain

$$h_1 \ \leq \ \frac{1}{2} \underbrace{h_0}_{< \ 2} \ h_0 \ < \ h_0 \ ,$$

and an induction argument shows

$$h_{k+1} \;<\; h_k \;<\; 2 \quad , \quad k \in \mathbb{N}_0 \; .$$

Moreover,

$$\|F(x^{k+1})\| \;<\; \|F(x^k)\| \;<\; \frac{2}{\omega} \quad \text{and} \quad \lim_{k\to\infty} \|F(x^k)\| \;=\; 0 \; ,$$

which implies

$$x^k \in \mathcal{L}_\omega \subset D \quad , \quad k \in \mathbb{N} \; .$$

Since $\mathcal{L}_\omega$ is bounded, there exist $x^* \in \overline{\mathcal{L}}_\omega$ and a subsequence $\mathbb{N}' \subset \mathbb{N}$ such that $x^k \to x^*$ $(k \in \mathbb{N}')$ and $F(x^*) = 0$.

•

Affine conjugacy

Assume that $D \subset \mathbb{R}^n$ is a convex set and that $f : D \to \mathbb{R}$ is a **strictly convex** functional. Consider the **minimization problem**

$$\min_{x\in D} f(x) \; .$$

Then, a **necessary and sufficient optimality condition** is given by the nonlinear equation

$$F(x) \;=\; \mathbf{grad}\; f(x) \;=\; f'(x)^T \;=\; 0 \quad , \quad x \in D \; .$$

We note that the Jacobian $F'(x) = f''(x)$ is **symmetric** and **uniformly positive definite** on D. In particular, $F'(x)^{1/2}$ is well defined and symmetric, positive definite as well.
Consequently, the **energy product**

$$(u, v)_E \;:=\; u^T F'(x) v \quad , \quad u, v, x \in D$$

defines locally an inner product with associated norm

$$\|u\|_E^2 \;=\; u^T F'(x) u \;=\; \|F'(x)^{1/2} u\|^2$$

which is referred to as a **local energy norm**.

For regular $B \in \mathbb{R}^{n\times n}$, we consider the **transformed minimization problem**

$$\min_y g(y) \quad , \quad g(y) \;:=\; f(By) \quad , \quad x \;=\; By \; .$$

We obtain the **optimality condition**

$$G(y) \;=\; \mathbf{grad}\; g(y) \;=\; (f'(By)B)^T \;=\; B^T f'(x)^T \;=\; B^T F(By) \;=\; 0$$

with the **transformed Jacobian**

$$G'(y) \;=\; B^T F'(x) B \;.$$

Hence, the **Jacobian transformation** is conjugate which motivates the notion of **affine conjugacy**.

An appropriate **affine conjugate Lipschitz condition** is as follows

$$\|F'(z)^{-1/2}\Big(F'(y)-F'(x)\Big)(y-x)\| \;\leq\; \omega\,\|F'(z)^{1/2}(y-x)\|^2\;.$$

2.3 Affine Conjugate Newton Convergence Theorem

Theorem 2.4 Affine Conjugate Newton-Mysovskikh Theorem

Assume that $D\subset\mathbb{R}^n$ is a convex domain and $f:D\to\mathbb{R}$ a strictly convex, twice continuously differentiable functional. Let $F(x)=f'(x)^T$ and $F'(x)=f''(x)$. Consider the minimization problem

$$\min_{x\in D} f(x) \tag{1.28}$$

and the associated optimality condition

$$F(x) \;=\; \mathbf{grad}\; f(x) \;=\; 0 \quad,\quad x\in D\;. \tag{1.29}$$

Note that (1.28) has a unique solution $x^*\in D$.
Let $x^0\in D$ be an initial guess and assume that the following conditions are satisfied:

$$\|F'(z)^{-1/2}\Big(F'(y)-F'(x)\Big)(y-x)\| \;\leq\; \omega\,\|F'(z)^{1/2}(y-x)\|^2 \tag{1.30}$$

for collinear $x,y,z\in D$ and

$$h_0 \;:=\; \omega\,\|F'(x^0)^{1/2}\Delta x^0\| \;<\; 2\;, \tag{1.31}$$

$$\mathcal{L}_0 \;:=\; \{x\in D\mid f(x)<f(x^0)\} \quad\text{is compact}\;. \tag{1.32}$$

Then, for the Newton iterates $x^k, k\in\mathbb{N}_0$, there holds:

(i) $x^k\in\mathcal{L}_0, k\in\mathbb{N}_0$, and $x^k\to x^*$ $(k\to\infty)$ with

$$\|F'(x^{k+1})^{1/2}\Delta x^{k+1}\| \;\leq\; \frac{1}{2}\,\omega\,\|F'(x^k)^{1/2}\Delta x^k\|^2\;. \tag{1.33}$$

(ii) For $\varepsilon_k := \|F'(x^k)^{1/2}\Delta x^k\|^2$ and the Kantorovich quantities $h_k := \omega\,\varepsilon_k^{1/2}$ we have

$$-\frac{1}{6}h_k\,\varepsilon_k \;\leq\; f(x^k)-f(x^{k+1}) \;-\; \frac{1}{2}\,\varepsilon_k \;\leq\; \frac{1}{6}h_k\,\varepsilon_k\;, \tag{1.34}$$

$$\frac{1}{6}\,\varepsilon_k \;\leq\; f(x^k)-f(x^{k+1}) \;\leq\; \frac{5}{6}\,\varepsilon_k\;. \tag{1.35}$$

(iii) We have the a priori estimate

$$f(x^0) \;-\; f(x^*) \;\leq\; \frac{\frac{5}{6}\,\varepsilon_0}{1-\frac{h_0}{2}}\;. \tag{1.36}$$

Proof: Assertion **(i)** and (1.33) can be verified as in the proof of the affine contravariant version of the Newton-Mysovskikh theorem.

For the proof of (1.34) in **(ii)**, observing

$$F'(x^k)\Delta x^k \;=\; -F(x^k)\;,$$

we obtain

$$\begin{aligned}
&f(x^{k+1}) - f(x^k) \;+\; \frac{1}{2}\,\|F'(x^k)^{1/2}\Delta x^k\|^2 \;= \\
&= \int_{s=0}^{1} < F(x^k + s\Delta x^k), \Delta x^k > \; ds \;-\; < F(x^k), \Delta x^k > \;- \\
&- \; < F'(x^k)\Delta x^k, \Delta x^k > \;+\frac{1}{2}\; < F'(x^k)\Delta x^k, \Delta x^k > = \\
&= \int_{s=0}^{1} < F(x^k + s\Delta x^k) - F(x^k), \Delta x^k > \; ds \;-\; \frac{1}{2}\; < F'(x^k)\Delta x^k, \Delta x^k > = \\
&= \int_{s=0}^{1}\int_{t=0}^{1} < F'(x^k + st\Delta x^k)\Delta x^k, \Delta x^k > \; dt\; ds - \frac{1}{2} < F'(x^k)\Delta x^k, \Delta x^k >= \\
&= \int_{s=0}^{1} s \int_{t=0}^{1} < \underbrace{(F'(x^k + st\Delta x^k) - F'(x^k))\Delta x_k}_{=:\; w_k}, \Delta x_k > \; dt\; ds \;= \\
&= \int_{s=0}^{1} s \int_{t=0}^{1} < F'(x^k)^{-1/2} w_k, F'(x^k)^{1/2}\Delta x^k > \; dt\; ds \;\leq \\
&\leq \int_{s=0}^{1} s \int_{t=0}^{1} \underbrace{\|F'(x^k)^{-1/2} w_k\|}_{\leq\; \omega st\|F'(x^k)^{1/2}\Delta x^k\|^2} \; \|F'(x^k)^{1/2}\Delta x^k\|\; dt\; ds \;\leq \\
&\leq \underbrace{\|F'(x^k)^{1/2}\Delta x^k\|^2}_{=\;\varepsilon_k}\; \underbrace{\omega\;\|F'(x^k)^{1/2}\Delta x^k\|}_{=\;h_k} \int_{s=0}^{1} s^2 \int_{t=0}^{1} t\; dt \;\leq\; \frac{1}{6}\; h_k\; \varepsilon_k\;,
\end{aligned}$$

which proves (1.34).

Using the right-hand side of (1.34) and $h_k < 2$ yields

$$f(x^k) \;-\; f(x^{k+1}) \;\leq\; (\frac{1}{2} \;+\; \frac{1}{6}h_k)\;\varepsilon_k \;<\; \frac{5}{6}\;\varepsilon_k\;.$$

Likewise, using the left-hand side of (1.34) and $h_k < 2$

$$f(x^k) \; - \; f(x^{k+1}) \; \geq \; (\frac{1}{2} \; - \; \frac{1}{6} h_k) \; \varepsilon_k \; > \; \frac{1}{6} \; \varepsilon_k \; .$$

Together, this proves (1.35).

In order to prove **(iii)**, we use (1.34) and obtain

$$0 \; \leq \; \omega^2 \; (f(x^0) - f(x^*)) \; \leq \; \omega^2 \; \sum_{k=0}^{\infty} (f(x^k) - f(x^{k+1}) \; < \; \frac{5}{6} \; \omega^2 \; \varepsilon_k \; =$$

$$= \; \frac{5}{6} \; h_k^2 \; = \; \frac{5}{6} \; 4 \; (\frac{1}{2} \; h_k)^2 \; .$$

Using

$$\frac{1}{2} \; h_{k+1} \; \leq \; (\frac{1}{2} \; h_k)^2 \; \leq \; \frac{1}{2} \; h_k \; < \; 1 \; ,$$

we further get

$$(\frac{1}{2} \; h_0)^2 \; + \; (\frac{1}{2} \; h_1)^2 \; + \; ... \; \leq$$

$$\leq \; (\frac{1}{2} \; h_0)^2 \; + \; (\frac{1}{2} \; h_0)^4 \; + \; (\frac{1}{2} \; h_1)^4 \; + \; ... \; \leq$$

$$\leq \; \frac{1}{4} \; h_0^2 \; \sum_{k=0}^{\infty} (\frac{1}{2} \; h_0)^k \; = \; \frac{\frac{1}{4} \; h_0^2}{1 - \frac{h_0}{2}} \; ,$$

which proves (1.36).

•

3. Inexact Newton Methods

We recall that Newton's method computes iterates successively as the solution of linear algebraic systems

$$\begin{aligned} F'(x^k)\,\Delta x^k &= -F(x^k) \quad , \quad k \in \mathbb{N}_0 \, , \\ x^{k+1} &= x^k + \Delta x^k \, . \end{aligned} \tag{1.37}$$

The classical convergence theorems of Newton-Kantorovich and Newton-Mysovskikh and its affine covariant, affine contravariant, and affine conjugate versions assume the **exact solution** of (1.37).
In practice however, in particular if the dimension is large, (1.37) will be solved by an **iterative method**. In this case, we end up with an **outer/inner iteration**, where the outer iterations are the Newton steps and the inner iterations result from the application of an iterative scheme to (1.37). It is important to tune the outer and inner iterations and to keep track of the iteration errors.
With regard to affine covariance, affine contravariance, and affine conjugacy the iterative scheme for the inner iterations has to be chosen in such a way, that it easily provides information about the

- **error norm** in case of **affine covariance**,
- **residual norm** in case of **affine contravariance**, and
- **energy norm** in case of **affine conjugacy**.

Except for **convex optimization**, we cannot expect $F'(x), x \in D$, to be symmetric positive definite. Hence, for affine covariance and affine contravariance we have to pick iterative solvers that are designed for nonsymmetric matrices. Appropriate candidates are

- **CGNE (Conjugate Gradient for the Normal Equations)** in case of **affine covariance**,
- **GMRES (Generalized Minimum RESidual)** in case of **affine contravariance**, and
- **PCG (Preconditioned Conjugate Gradient)** in case of **affine conjugacy**.

3.1 Affine Covariant Inexact Newton Methods

3.1.1 CGNE (Conjugate Gradient for the Normal Equations)

We assume $A \in \mathbb{R}^{n\times n}$ to be a regular, nonsymmetric matrix and $b \in \mathbb{R}^n$ to be given and look for $y^* \in \mathbb{R}^n$ as the unique solution of the **linear algebraic system**

$$Ay \;=\; b\,. \tag{1.38}$$

As the name already suggests, **CGNE** is the conjugate gradient method applied to the **normal equations**:
It solves the system

$$AA^Tz \;=\; b\,, \tag{1.39}$$

for z and then computes y according to

$$y \;=\; A^Tz\,. \tag{1.40}$$

The implementation of **CGNE** is as follows:

CGNE Initialization:

Given an **initial guess** $y_0 \in \mathbb{R}^n$, compute the **residual** $r_0 = b - Ay_0$ and set

$$\begin{aligned} p_0 \;&=\; r_0 \quad,\quad p_0 \;=\; 0\,,\\ \beta_0 \;&=\; 0 \quad,\quad \sigma_0 \;=\; \|r_0\|^2\,. \end{aligned}$$

CGNE Iteration Loop: For $1 \le i \le i_{max}$ compute

$$p_i \;=\; A^Tr_{i-1} \;+\; \beta_{i-1}\,p_{i-1} \quad,\quad \alpha_i \;=\; \frac{\sigma_{i-1}}{\|p_i\|^2}\,,$$

$$y_i \;=\; y_{i-1} \;-\; \alpha_i p_i \quad,\quad \gamma_{i-1}^2 \;=\; \alpha_i\sigma_{i-1}\,,$$

$$r_i \;=\; r_{i-1} \;-\; \alpha_i Ap_i \quad,\quad \sigma_i \;=\; \|r_i\|^2\,,$$

$$\beta_i \;=\; \frac{\sigma_i}{\sigma_{i-1}}\,.$$

CGNE has the **error minimizing property**

$$\|y^* - y_i\| \;=\; \min_{v\in\mathcal{K}_i(A^Tr_0,A^TA)} \|y^* - v\|\,, \tag{1.41}$$

where $\mathcal{K}_i(A^Tr_0, A^TA)$ stands for the **Krylov subspace**

$$\mathcal{K}_i(A^Tr_0, A^TA) \;:=\; \mathrm{span}\{A^Tr_0, (A^TA)A^Tr_0, ..., (A^TA)^{i-1}A^Tr_0\}\,. \tag{1.42}$$

Lemma 3.1 Representation of the iteration error

Let $\varepsilon_i := \|y^* - y_i\|^2$ be the square of the **CGNE iteration error** with respect to the i-th iterate. Then, there holds

$$\varepsilon_i \;=\; \sum_{j=i}^{n-1} \gamma_j^2 \;. \tag{1.43}$$

Proof. CGNE has the **Galerkin orthogonality**

$$(y_i - y_0, y_{i+m} - y_i) \;=\; 0 \quad , \quad m \in \mathbb{N} \;. \tag{1.44}$$

Setting $m = 1$, this implies the orthogonal decomposition

$$\|y_{i+1} - y_0\|^2 \;=\; \|y_{i+1} - y_i\|^2 \;+\; \|y_i - y_0\|^2 \;, \tag{1.45}$$

which readily gives

$$\|y_i - y_0\|^2 \;=\; \sum_{j=0}^{i-1} \|y_{j+1} - y_j\|^2 \;=\; \sum_{j=0}^{i-1} \gamma_j^2 \;. \tag{1.46}$$

On the other hand, observing $y_n = y^*$, for $m = n - i$ the Galerkin orthogonality yields

$$\underbrace{\|y^* - y_0\|^2}_{= \sum_{j=0}^{n-1} \gamma_j^2} \;=\; \underbrace{\|y^* - y_i\|^2}_{= \varepsilon_i^2} \;+\; \underbrace{\|y_i - y_0\|^2}_{= \sum_{j=0}^{i-1} \gamma_j^2} \;. \tag{1.47}$$

Computable lower bound for the iteration error

It follows readily from Lemma 3.1 that the computable quantity

$$[\varepsilon_i] \;:=\; \sum_{j=i}^{i+m} \gamma_j^2 \quad , \quad m \in \mathbb{N}, \tag{1.48}$$

provides a **lower bound** for the iteration error.
In practice, we will test the **relative error norm** according to

$$\delta_i \;:=\; \frac{\|y^* - y_i\|}{\|y_i\|} \;\approx\; \frac{\sqrt{[\varepsilon_i]}}{\|y_i\|} \;\leq\; \overline{\delta} \;, \tag{1.49}$$

where $\overline{\delta}$ is a **user specified accuracy**.

3.1.2 Convergence of affine covariant inexact Newton methods

We denote by $\delta x^k \in \mathbb{R}^n$ the result of an inner iteration, e.g., CGNE, for the solution of (1.37). Then, it is easy to see that the iteration error $\delta x^k - \Delta x^k$ satisfies the **error equation**

$$F'(x^k)(\delta x^k - \Delta x^k) \;=\; F(x^k) \;+\; F'(x^k)\delta x^k \;=:\; r^k \;. \tag{1.50}$$

We will measure the impact of the inexact solution of (1.37) by the **relative error**

$$\delta_k \;:=\; \frac{\|\delta x^k - \Delta x^k\|}{\|\delta x^k\|} \;. \tag{1.51}$$

Theorem 3.1 Affine covariant convergence theorem for the inexact Newton method. Part I: Linear convergence

Suppose that that $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$ is continuously differentiable on D with invertible Fréchet derivatives $F'(x), x \in \mathbb{R}^n$. Assume further that the following **affine covariant Lipschitz condition** is satisfied

$$\|F'(z)^{-1}\Big(F'(y) - F'(x)\Big)v\| \;\leq\; \omega\, \|y - x\|\, \|v\| \;, \tag{1.52}$$

where $x, y, z \in D, v \in \mathbb{R}^n$.
Assume that $x^0 \in D$ is an **initial guess** for the **outer Newton iteration** and that $\delta x^0 = 0$ is chosen as the startiterate for the **inner iteration**. Consider the **Kantorovich quantities**

$$h_k \;:=\; \omega\, \|\Delta x^k\| \quad , \quad h_k^\delta \;:=\; \omega\, \|\delta x^k\| \;=\; \frac{h_k}{\sqrt{1+\delta_k^2}} \tag{1.53}$$

associated with the outer and inner iteration.
Assume that

$$h_0 \;<\; 2\,\overline{\Theta} \quad , \quad 0 \;\leq\; \overline{\Theta} \;<\; 1 \;, \tag{1.54}$$

and **control the inner iterations** according to

$$\vartheta(h_k, \delta_k) \;:=\; \frac{\frac{1}{2}h_k^\delta \;+\; \delta_k(1 + h_k^\delta)}{\sqrt{1 \;+\; \delta_k^2}} \;\leq\; \overline{\Theta} \;<\; 1 \;, \tag{1.55}$$

which implies **linear convergence**.
Note that a **necessary condition** for $\vartheta(h_k, \delta_k) \leq \overline{\Theta}$ is that it holds true for $\delta_k = 0$, which is satisfied due to assumption (1.37).

Then, there holds:

(i) The **Newton CGNE iterates** $x^k, k \in \mathbb{N}_0$ stay in

$$\overline{B}(x^0, \rho) \quad , \quad \rho := \frac{\|\delta x^0\|}{1 - \overline{\Theta}} \tag{1.56}$$

and **converge linearly** to some $x^* \in \overline{B}(x^0, \rho)$ with $F(x^*) = 0$.

(ii) The **exact Newton increments** decrease monotonically according to

$$\frac{\|\Delta x^{k+1}\|}{\|\Delta x^k\|} \leq \overline{\Theta} \,, \tag{1.57}$$

whereas for the **inexact Newton increments** we have

$$\frac{\|\delta x^{k+1}\|}{\|\delta x^k\|} \leq \frac{\sqrt{1+\delta_k^2}}{\sqrt{1+\delta_{k+1}^2}}\,\overline{\Theta} \leq \overline{\Theta} \,. \tag{1.58}$$

Proof. By elementary calculations we find

$$\|\Delta x^{k+1}\| = \|F'(x^{k+1})^{-1} F(x^{k+1})\| = \tag{1.59}$$

$$= \|F'(x^{k+1})^{-1}\Big[F(x^{k+1}) - F(x^k)\Big] + F'(x^{k+1})^{-1} \underbrace{F(x^k)}_{= r^k - F'(x^k)\delta x^k} \| \leq$$

$$= \|F'(x^{k+1})^{-1}\Big[F(x^{k+1}) - F(x^k) - F'(x^k)\delta x^k\Big]\| +$$

$$+ \|F'(x^{k+1})^{-1} \underbrace{r^k}_{= F'(x^k)(\delta x^k - \Delta x^k)} \| \leq$$

$$\leq \underbrace{\int_0^1 \|F'(x^{k+1})^{-1}\Big[F'(x^k + t\delta x^k) - F'(x^k)\Big]\delta x^k\| dt}_{=: \mathbf{I}} +$$

$$+ \underbrace{\|F'(x^{k+1})^{-1} F'(x^k)(\delta x^k - \Delta x^k)\|}_{=: \mathbf{II}} \,.$$

Using the **affine covariant Lipschitz condition** (1.52), the first term on the right-hand side in (1.59) can be estimated according to

$$\mathbf{I} \;\leq\; \omega\,\|\delta x^k\|^2 \int_0^1 t\;dt \;=\; \frac{1}{2}\,\omega\,\|\delta x^k\|^2\;. \tag{1.60}$$

For the second term we obtain by the same argument

$$\begin{aligned}
\mathbf{II} \;&=\; \|F'(x^{k+1})^{-1}\Big[F'(x^k)(\delta x^k-\Delta x^k)\;\pm\;F'(x^{k+1})(\delta x^k-\Delta x^k)\Big]\| \;\leq \qquad (1.61)\\
&\leq\; \|F'(x^{k+1})^{-1}(F'(x^{k+1})-F'(x^k))(\delta x^k-\Delta x^k)\| \;+\\
&\quad+\;\|F'(x^{k+1})^{-1}F'(x^{k+1})(\delta x^k-\Delta x^k)\| \;\leq\\
&\leq\; \frac{1}{2}\,\omega\|\delta x^k\|\;\|\delta x^k-\Delta x^k\| \;+\; \|\delta x^k-\Delta x^k\|^2\;.
\end{aligned}$$

Combining (1.60) and (1.61) yields

$$\begin{aligned}
\frac{\|\delta x^{k+1}\|}{\|\delta x^k\|} \;&\leq\; \frac{1}{2}\underbrace{\omega\,\|\delta x^k\|}_{=\,h_k^\delta} + \frac{1}{2}\underbrace{\omega\,\|\delta x^k\|\,\frac{\|\delta x^k-\Delta x^k\|}{\|\delta x^k}}_{=\,\delta_k\;h_k^\delta} + \underbrace{\frac{\|\delta x^k-\Delta x^k\|}{\|\delta x^k\|}}_{=\,\delta_k} \;\leq\\
&\leq\; h_k^\delta \;+\; \delta_k\,(1+h_k^\delta)\;.
\end{aligned}$$

Observing (1.53), we finally get

$$\frac{\|\Delta x^{k+1}\|}{\|\Delta x^k\|} \;\leq\; \vartheta(h_k,\delta_k) \;=\; \frac{\frac{1}{2}h_k^\delta \;+\; \delta_k(1+h_k^\delta)}{\sqrt{1\;+\;\delta_k^2}} \;\leq\; \overline{\Theta} \;<\; 1\;, \tag{1.62}$$

which implies **linear convergence**.
Note that a necessary condition for $\vartheta(h_k,\delta_k)\leq\overline{\Theta}$ is that it holds true for $\delta_k=0$, which is satisfied due to assumption (1.54).
For the **contraction of the inexact Newton increments** we get

$$\frac{\|\delta x^{k+1}\|}{\|\delta x^k\|} \;=\; \sqrt{\frac{1+\delta_k^2}{1+\delta_{k+1}^2}}\;\frac{\|\Delta x^{k+1}\|}{\|\Delta x^k\|} \;\leq\; \sqrt{\frac{1+\delta_k^2}{1+\delta_{k+1}^2}}\;\overline{\Theta} \;\leq\; \overline{\Theta}\;. \tag{1.63}$$

It can be easily shown that $\{x^k\}_{\mathbb{N}_0}$ is a **Cauchy sequence** in $\overline{B}(x^0,\rho)$. Consequently, there exists $x^*\in\overline{B}(x^0,\rho)$ such that $x^k\to x^*$ $(k\to\infty)$. Since

$$\underbrace{F'(x^k)\delta x^k}_{\to\,0} \;=\; \underbrace{F(x^k)\;+\;r^k}_{\to\,F(x^*)}\;,$$

we conclude $F(x^*)=0$. •

Theorem 3.2 Affine covariant convergence theorem for the inexact Newton method. Part II: Quadratic convergence

Under the same assumptions on $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$ as in Theorem 3.1 suppose that the initial guess $x^0 \in D$ satisfies

$$h_0 \;<\; \frac{2}{1+\rho} \tag{1.64}$$

for some appropriate $\rho > 0$ and **control the inner iterations** such that

$$\delta_k \;\leq\; \frac{\rho}{2}\,\frac{h_k^\delta}{1+h_k^\delta}\;. \tag{1.65}$$

Then, there holds:

(i) The **Newton CGNE iterates** $x^k, k \in \mathbb{N}_0$ stay in

$$\overline{B}(x^0,\overline{\rho}) \quad , \quad \overline{\rho} \;:=\; \frac{\|\delta x^0\|}{1-\frac{1+\rho}{2}\,h_0} \tag{1.66}$$

and **converge quadratically** to some $x^* \in \overline{B}(x^0,\overline{\rho})$ with $F(x^*) = 0$.

(ii) The **exact Newton increments** and the **inexact Newton increments** decrease quadratically according to

$$\|\Delta x^{k+1}\| \;\leq\; \frac{1+\rho}{2}\,\omega\,\|\Delta x^k\|^2\;, \tag{1.67}$$

$$\|\delta x^{k+1}\| \;\leq\; \frac{1+\rho}{2}\,\omega\,\|\delta x^k\|^2\;. \tag{1.68}$$

Proof. We proceed as in the proof of Theorem 3.1 to obtain

$$\frac{\|\Delta x^{k+1}\|}{\|\Delta x^k\|} \;\leq\; \vartheta(h_k,\delta_k) \;=\; \frac{\frac{1}{2}h_k^\delta \;+\; \delta_k(1+h_k^\delta)}{\sqrt{1\;+\;\delta_k^2}}\;.$$

and

$$\frac{\|\delta x^{k+1}\|}{\|\delta x^k\|} \;=\; \sqrt{\frac{1+\delta_k^2}{1+\delta_{k+1}^2}}\;\frac{\|\Delta x^{k+1}\|}{\|\Delta x^k\|}\;.$$

In view of (1.65) we get the further estimates

$$\frac{\|\Delta x^{k+1}\|}{\|\Delta x^k\|} \;\leq\; \frac{1+\rho}{2}\,\frac{h_k}{1+\delta_k^2} \;\leq\; \frac{1+\rho}{2}\,h_k\;.$$

and

$$\frac{\|\delta x^{k+1}\|}{\|\delta x^k\|} \;\leq\; \frac{1+\rho}{2}\,\frac{h_k^\delta}{\sqrt{1+\delta_{k+1}^2}} \;\leq\; \frac{1+\rho}{2}\,h_k^\delta\;,$$

from which (1.67) and (1.68) follow by the definition of the Kantorovich quantities.
In order to deduce quadratic convergence we have to make sure that the initial increments ($k=0$) are small enough, i.e.,

$$\frac{1+\rho}{2}\,h_0^\delta \;\leq\; \frac{1+\rho}{2}\,h_0 \;<\; 1\;. \tag{1.69}$$

Furthermore, (1.68) and (1.69) allow us to show that the iterates $x^k, k\in\mathbb{N}$ stay in $\overline{B}(x^0,\overline{\rho})$. Indeed, (1.68) implies

$$\|\delta x^j\| \;\leq\; \frac{1+\rho}{2}\,h_{j-1}\,\|\delta x^{j-1}\| \;\leq\; \frac{1+\rho}{2}\,h_0\,\|\delta x^{j-1}\|\;,\; j\in\mathbb{N}\;,$$

and hence,

$$\|x^k-x^*\| \;\leq\; \sum_{j=0}^{k}\|\delta x^j\| \;\leq\; \|\delta x^0\|\sum_{j=0}^{k}(\frac{1+\rho}{2}\,h_0)^j \;\leq\; \frac{\|\delta x^0\|}{1-\frac{1+\rho}{2}\,h_0}\;.\bullet$$

3.1.3 Algorithmic aspects of affine covariant inexact Newton methods

(i) Convergence monitor

Let us assume that the quantity $\overline{\Theta}<1$ in both the **linear convergence mode** and the **quadratic convergence mode** has been specified and let us further assume that we use CGNE with $\delta x_0^k=0$ in the inner iteration.
Then, (1.58) suggests the **monotonicity test**

$$\tilde{\Theta}_k \;:=\; \sqrt{\frac{1+\overline{\delta}_{k+1}^2}{1+\overline{\delta}_k^2}}\;\frac{\|\delta x^{k+1}\|}{\|\delta x^k\|} \;\leq\; \overline{\Theta}\;, \tag{1.70}$$

where $\overline{\delta}_k^2$ and $\overline{\delta}_{k+1}^2$ are computationally available estimates of δ_k^2 and δ_{k+1}^2.

(ii) Termination criterion

We recall that the termination criterion for the exact Newton iteration with respect to a user specified accuracy $XTOL$ is given by

$$\frac{\|\Delta x^k\|}{1-\Theta_{k-1}^2} \;\leq\; \mathrm{XTOL}\;.$$

According to (1.53) we have

$$\|\Delta x^k\| \;=\; \sqrt{1+\delta_k^2}\;\|\delta x^k\|.$$

Consequently, replacing Θ_{k-1} and δ_k by the computable quantities $\tilde{\Theta}_{k-1}$ and $\overline{\delta}_k$, we arrive at the termination criterion

$$\frac{\sqrt{1+\overline{\delta}_k^2}}{1-\tilde{\Theta}_{k-1}^2} \;\leq\; \text{XTOL}\;. \tag{1.71}$$

(iii) Balancing outer and inner iterations

According to (1.55) of Theorem 3.1, in the **linear convergence mode** the adaptive termination criterion for the inner iteration is

$$\vartheta(h_k,\delta_k) \;:=\; \frac{\frac{1}{2}h_k^\delta \;+\; \delta_k(1+h_k^\delta)}{\sqrt{1\;+\;\delta_k^2}} \;\leq\; \overline{\Theta} \;<\; 1\;.$$

On the other hand, in view of (1.65) of Theorem 3.2, in the **quadratic convergence mode** the termination criterion is

$$\delta_k \;\leq\; \frac{\rho}{2}\,\frac{h_k^\delta}{1+h_k^\delta}\;.$$

Since the **theoretical Kantorovich quantities** (cf. (1.53))

$$h_k^\delta \;=\; \omega\;\|\delta x^k\| \;=\; \frac{h_k}{\sqrt{1+\delta_k^2}}$$

are not directly accessible, we have to replace them by **computationally available estimates** $[h_k^\delta]$.
We recall that for h_k we have the a priori estimate

$$[h_k] \;=\; 2\;\Theta_{k-1}^2 \;\leq\; h_k\;.$$

Consequently, replacing δ_k by $\overline{\delta}_k$, h_k by $[h_k]$, and Θ_{k-1} by $\tilde{\Theta}_{k-1}$ (cf. (1.70)), we get the **a priori estimates**

$$[h_k^\delta] \;=\; \frac{[h_k]}{\sqrt{1+\overline{\delta}_k^2}} \quad,\quad [h_k] \;=\; 2\;\tilde{\Theta}_{k-1}^2 \quad,\quad k\in\mathbb{N}\;. \tag{1.72}$$

For $k=0$, we choose $\overline{\delta}_0=\delta_0=\frac{1}{4}$.
In practice, for $k\geq 1$ we begin with the quadratic convergence mode and switch

to the linear convergence mode as soon as the approximate contraction factor $\tilde{\Theta}_k$ is below some prespecified threshold value $\overline{\Theta} \leq \frac{1}{2}$.

(iii)$_1$ Quadratic convergence mode

The computationally realizable **termination criterion for the inner iteration in the quadratic convergence mode** is

$$\overline{\delta}_k \;\leq\; \frac{\rho}{2}\,\frac{[h_k^\delta]}{1+[h_k^\delta]}\;. \tag{1.73}$$

Inserting (1.72) into (1.73), we obtain a simple nonlinear equation in $\overline{\delta}_k$.

Remark 3.1 Validity of the approximate termination criterion

Observing that the right-hand side in (1.73) is a monotonically increasing function of $[h_k^\delta]$, and taking $[h_k^\delta] \leq h_k^\delta$ into account, it follows that for $\delta_k \leq \overline{\delta}_k$ the approximate termination criterion (1.73) implies the exact termination criterion (1.65).

Remark 3.2 Computational work in the quadratic convergence mode

Since $\overline{\delta}_k \to 0$ $(k \to \infty)$ is enforced, it follows that:
The more the iterates x^k approach the solution x^*, the more computational work is required for the inner iterations to guarantee quadratic convergence of the outer iteration.

(iii)$_2$ Linear convergence mode

We switch to the linear convergence mode, once the criterion

$$\tilde{\Theta}_k \;<\; \overline{\Theta} \tag{1.74}$$

is met.
The computationally realizable **termination criterion for the inner iteration in the linear convergence mode** is

$$[\vartheta(h_k,\delta_k)] \;:=\; \vartheta([h_k],\overline{\delta}_k) \;=\; \frac{\frac{1}{2}[h_k^\delta] \;+\; \overline{\delta}_k(1+[h_k^\delta])}{\sqrt{1 \;+\; \overline{\delta}_k^2}} \;\leq\; \overline{\Theta}\;. \tag{1.75}$$

Remark 3.3 Validity of the approximate termination criterion

Since the right-hand side in (1.75) is a monotonically increasing function in $[h_k^\delta]$ and $[h_k^\delta] \leq h_k^\delta$, the estimate provided by (1.75) may be too small and thus result in an **overestimation** of δ_k. However, since the exact quantities and their a priori estimates both tend to zero as k approaches infinity, asymptotically we may rely on (1.75).

In practice, we require the **monotonicity test** (1.70) in CGNE and run the inner iterations until $\overline{\delta}_k$ satisfies (1.75) or divergence occurs, i.e.,

$$\tilde{\Theta}_k \; > \; 2\,\overline{\Theta}\;.$$

Remark 3.4 Computational work in the linear convergence mode

As opposed to the quadratic convergence mode, we observe

The more the iterates x^k approach the solution x^*, the less computational work is required for the inner iterations to guarantee linear convergence of the outer iteration.

3.2 Affine Contravariant Inexact Newton Methods

3.2.1 GMRES (Generalized Minimum RESidual)

The **Generalized Minimum RESidual Method (GMRES** is an iterative solver for nonsymmetric linear algebraic systems which generates an **orthogonal basis** of the **Krylov subspace**

$$\mathcal{K}_i(r_0, A) \;:=\; \text{span}\{r_0, Ar_0, ..., A^{i-1}r_0\}\;. \tag{1.76}$$

by a modified Gram-Schmidt orthogonalization called the **Arnoldi method**. The inner product coefficients are stored in an **upper Hessenberg matrix** so that an approximate solution can be obtained by the solution of a **least-squares problem** in terms of that Hessenberg matrix:

GMRES Initialization:

Given an **initial guess** $y_0 \in \mathbb{R}^n$, compute the **residual** $r_0 = b - Ay_0$ and set

$$\beta \;:=\; \|r_0\| \quad , \quad v_1 \;:=\; \frac{r_0}{\beta} \quad , \quad V_1 \;:=\; v_1\;. \tag{1.77}$$

GMRES Iteration Loop: For $1 \le i \le i_{max}$:

I. Orthogonalization:

$$\hat{v}_{i+1} \;=\; Av_i \;-\; V_i h_i\;, \tag{1.78}$$

$$\text{where} \quad h_i \;=\; V_i^T A v_i\;. \tag{1.79}$$

II. Normalization:

$$\hat{v}_{i+1} \;=\; \frac{\hat{v}_{i+1}}{\|\hat{v}_{i+1}\|}\;. \tag{1.80}$$

III. Update:

$$V_{i+1} \;=\; \left(V_i \; v_{i+1}\right)\;. \tag{1.81}$$

$$H_i \;=\; \begin{pmatrix} h_i \\ \|\hat{v}_{i+1}\| \end{pmatrix} \quad , \quad i = 1\;, \tag{1.82}$$

$$H_i \;=\; \begin{pmatrix} H_{i-1} & h_i \\ 0 & \|\hat{v}_{i+1}\| \end{pmatrix} \quad , \quad i > 1\;. \tag{1.83}$$

IV. Least squares problem: Compute z_i as the solution of

$$\|\beta\, e_1 \;-\; H_i z_i\| \;=\; \min_{z\in\mathbb{R}^n} \|\beta\, e_1 \;-\; H_i z\| \;. \tag{1.84}$$

V. Approximate solution:

$$y_i \;=\; V_i z_i \;+\; y_0 \;. \tag{1.85}$$

GMRES has the **residual norm minimizing property**

$$\|b \;-\; A y_i\| \;=\; \min_{z\in\mathcal{K}_i(r_0,A)} \|b \;-\; A z\| \;. \tag{1.86}$$

Moreover, the **inner residuals decrease monotonically**

$$\|r_{i+1}\| \;\leq\; \|r_i\| \quad , \quad i \in \mathbb{N}_0 \;. \tag{1.87}$$

Termination criterion for the GMRES iteration

The residuals satisfy the **orthogonality relation**

$$(r_i, r_i \;-\; r_0) \;=\; 0 \quad , \quad i \in \mathbb{N} \;, \tag{1.88}$$

from which we readily deduce

$$\|r_0\|^2 \;=\; \|r_i - r_0\|^2 \;+\; \|r_i\|^2 \quad , \quad i \in \mathbb{N} \;. \tag{1.89}$$

We define the **relative residual norm error**

$$\eta_i \;:=\; \frac{\|r_i\|}{\|r_0\|} \;. \tag{1.90}$$

Clearly, $\eta_i < 1, i \in \mathbb{N}$, and

$$\eta_{i+1} \;<\; \eta_i \quad \text{if } \eta_i \neq 0 \;. \tag{1.91}$$

Consequently, given a **user specified accuracy** $\overline{\eta}$, an appropriate **adaptive termination criterion** is

$$\eta_i \;\leq\; \overline{\eta} \;. \tag{1.92}$$

We note that, in terms of η_i, (1.89) can be written as

$$\|r_i - r_0\|^2 \;=\; (1 - \eta_i^2)\; \|r_0\|^2 \;. \tag{1.93}$$

3.2.2 Convergence of affine contravariant inexact Newton methods

We denote by $\delta x^k \in \mathbb{R}^n$ the result of the inner GMRES iteration. As **initial values** for GMRES we choose

$$\delta x_0^k \;=\; 0 \quad , \quad r_0^k \;=\; F(x^k) \; . \tag{1.94}$$

Consequently, during the inner GMRES iteration the **relative error** $\eta_i, i \in \mathbb{N}_0$, in the residuals satisfies

$$\eta_i \;=\; \frac{\|r_i^k\|}{\|F(x^k)\|} \;\leq\; 1 \quad , \quad \eta_{i+1} \;<\; \eta_i \; , \text{ if } \eta_i \neq 0 \; . \tag{1.95}$$

In the sequel, we drop the subindices i for the inner iterations and refer to η_k as the final value of the inner iterations at each outer iteration step k.

Theorem 3.3 Affine contravariant convergence theorem for the inexact Newton GMRES method. Part I: Linear convergence

Suppose that $F : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$ is continuously differentiable on D and let $x^0 \in D$ be some initial guess. Let further the following **affine contravariant Lipschitz condition** be satisfied

$$\|(F'(y) - F'(x))(y-x)\| \;\leq\; \omega \, \|F'(x)(y-x)\|^2 \; , \; x, y \in D \; , \; \omega \geq 0 \; . \tag{1.96}$$

Assume further that the level set

$$\mathcal{L}_0 \;:=\; \{x \in \mathbb{R}^n \mid \|F(x)\| \;\leq\; \|F(x^0)\|\} \tag{1.97}$$

is a **compact subset** of D.
In terms of the **Kantorovich quantities**

$$h_k \;:=\; \omega \, \|F(x^k)\| \quad , \quad k \in \mathbb{N}_0 \; . \tag{1.98}$$

the **outer residual norms** can be bounded according to

$$\|F(x^{k+1})\| \;\leq\; \left(\eta_k + \frac{1}{2} \, (1 - \eta_k^2) \, h_k\right) \|F(x^k)\| \; . \tag{1.99}$$

Assume that

$$h_0 \;<\; 2 \tag{1.100}$$

and **control the inner iterations** according to

$$\eta_k \;\leq\; \overline{\Theta} \;-\; \frac{1}{2} \, h_k \; , \tag{1.101}$$

for some $\frac{h_0}{2} < \overline{\Theta} < 1$.
Then, the **Newton GMRES iterates** $x^k, k \in \mathbb{N}_0$ stay in $\mathcal{L}_0$ and **converge linearly** to some $x^* \in \mathcal{L}_0$ with $F(x^*) = 0$ at an estimated rate

$$\|F(x^{k+1})\| \;\leq\; \overline{\Theta}\, \|F(x^k)\| \;. \tag{1.102}$$

Proof. We recall that the Newton GMRES iterates satisfy

$$F'(x^k)\; \delta x^k \;=\; -\, F(x^k) \;+\; r^k \;, \tag{1.103}$$

$$x^{k+1} \;=\; x^k \;+\; \delta x^k \;. \tag{1.104}$$

It follows from the generalized mean value theorem that

$$F(x^{k+1}) \;=\; F(x^k) \;+\; \int_0^1 F'(x^k + t\delta x^k)\; \delta x^k \; dt \;. \tag{1.105}$$

Consequently, replacing $F(x^k)$ in (1.105) by (1.103), we obtain

$$\begin{aligned}
\|F(x^{k+1})\| \;&=\; \| \int_0^1 \Big(F'(x^k + t\delta x^k) - F'(x^k) \Big)\; \delta x^k \; dt \;+\; r^k \| \;\leq \\
&\leq\; \int_0^1 \| \Big(F'(x^k + t\delta x^k) - F'(x^k) \Big)\; \delta x^k \| \; dt \;+\; \|r^k\| \;\leq \\
&\leq\; \frac{1}{2}\, \omega\, \|F'(x^k)\; \delta x^k\|^2 \;+\; \|r^k\| \;\leq \\
&\leq\; \frac{1}{2}\, \omega\, \|F(x^k) - r^k\|^2 \;+\; \|r^k\| \;.
\end{aligned}$$

We recall (1.93)

$$\|r^k - F(x^k)\|^2 \;=\; (1 - \eta_k^2)\; \|F(x^k)\|^2 \;,$$

from which (1.99) can be immediately deduced.
Now, in view of (1.101), (1.99) yields

$$\begin{aligned}
\|F(x^{k+1})\| \;&\leq\; \Big(\underbrace{\eta_k}_{\leq\, \overline{\Theta} - \frac{1}{2} h_k} + \frac{1}{2}\, (1 - \eta_k^2) h_k \Big)\; \|F(x^k)\| \;\leq \\
&\leq\; (\overline{\Theta} \;-\; \frac{1}{2}\, \eta_k^2\; h_k)\; \|F(x^k)\| \;\leq\; \overline{\Theta}\; \|F(x^k)\| \;.
\end{aligned}$$

Taking advantage of the previous inequality, by induction on k it follows that

$$x^k \in \mathcal{L}_0 \subset D \quad , \quad k \in \mathbb{N}_0 \;.$$

Hence, there exists a subsequence $\mathbb{N}' \subset \mathbb{N}$ and an $x^* \in \mathcal{L}_0$ such that $x^k \to x^*$ $(k \in \mathbb{N}' \to \infty)$ and $F(x^*) = 0$. Moreover, since

$$\begin{aligned} \|F(x^{k+\ell}) - F(x^k)\| \;&\le\; \|F(x^{k+\ell})\| \;+\; \|F(x^k)\| \;\le\; (1 + \overline{\Theta}^\ell)\, \|F(x^k)\| \;\le \\ &\le\; (1 + \overline{\Theta}^\ell)\, \overline{\Theta}^k\, \|F(x^0)\| \to 0 \quad (k \in \mathbb{N} \to \infty)\;, \end{aligned}$$

the whole sequence must converge to x^*.

•

Theorem 3.4 Affine contravariant convergence theorem for the inexact Newton GMRES method. Part II: Quadratic convergence

Under the same assumptions on $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$ as in Theorem 3.3 suppose that the initial guess $x^0 \in D$ satisfies

$$h_0 \;<\; \frac{2}{1+\rho} \tag{1.106}$$

for some appropriate $\rho > 0$ and **control the inner iterations** such that

$$\frac{\eta_k}{1-\eta_k^2} \;\le\; \frac{\rho}{2}\, h_k\;. \tag{1.107}$$

Then, the **Newton GMRES iterates** $x^k, k \in \mathbb{N}_0$ stay in $\mathcal{L}_0$ and **converge quadratically** to some $x^* \in \overline{B}(x^0, \overline{\rho})$ with $F(x^*) = 0$ at an estimated rate

$$\|F(x^{k+1})\| \;\le\; \frac{1}{2}\, \omega\, (1+\rho)\, (1-\eta_k^2)\, \|F(x^k)\|^2\;. \tag{1.108}$$

Proof. Inserting (1.107) into (1.99) and observing $h_k = \omega \|F(x^k)\|$ gives the assertion.

•

3.2.3 Algorithmic aspects of affine contravariant inexact Newton methods

(i) Convergence monitor

Throughout the inexact Newton GMRES iteration we use the **residual monotonicity test**

$$\tilde{\Theta}_k \;:=\; \frac{\|F(x^{k+1})\|}{\|F(x^k)\|} \;\le\; \overline{\Theta} \;<\; 1\;. \tag{1.109}$$

The iteration is considered as **divergent**, if

$$\tilde{\Theta}_k \;>\; \overline{\Theta}\;. \tag{1.110}$$

(ii) Termination criterion

As in the exact Newton iteration, specifying a **residual accuracy** $FTOL$, the termination criterion for the inexact Newton GMRES iteration is

$$\|F(x^k)\| \;\leq\; \text{FTOL} \;. \tag{1.111}$$

(iii) Balancing outer and inner iterations

With regard to (1.101) of Theorem 3.3, in the **linear convergence mode** the adaptive termination criterion for the inner GMRES iteration is

$$\eta_k \;\leq\; \overline{\Theta} \;-\; \frac{1}{2}\, h_k \;,$$

whereas, in view of (1.107) of Theorem 3.4, in the **quadratic convergence mode** the termination criterion is

$$\frac{\eta_k}{1-\eta_k^2} \;\leq\; \frac{\rho}{2}\, h_k \;.$$

Again, we replace the theoretical Kantorovich quantities h_k by some computationally easily available a priori estimates. We distinguish between the quadratic and the linear convergence mode:

(iii)$_1$ Quadratic convergence mode

We recall the termination criterion (1.107) for the quadratic convergence mode

$$\frac{\eta_k}{1-\eta_k^2} \;\leq\; \frac{\rho}{2}\, h_k \;.$$

It suggests the **a posteriori estimate**

$$[h_k]_2 \;:=\; \frac{2\;\Theta_k}{(1+\rho)\;(1-\eta_k^2)} \;\leq\; h_k \;.$$

In view of $h_{k+1} = \Theta_k h_k$, this implies the **a priori estimate**

$$[h_{k+1}] \;:=\; \Theta_k\;[h_k]_2 \;\leq\; \Theta_k\; h_k \;=\; h_{k+1} \;. \tag{1.112}$$

Using (1.112) in (1.107) results in the computationally feasible termination criterion

$$\frac{\eta_k}{1-\eta_k^2} \;\leq\; \frac{1}{2}\,\rho\;[h_k] \quad , \quad \rho \approx 1.0 \;. \tag{1.113}$$

(iii)$_2$ Linear convergence mode

We switch from the quadratic to the linear convergence mode, if the local contraction factor satisfies

$$\Theta_k \;<\; \overline{\Theta}\;. \tag{1.114}$$

The proof of the previous theorems reveals

$$\|F(x^{k+1}) - r^k\| \;\leq\; \frac{\omega}{2}\,\|F(x^k) - r^k\|^2 \;=\; \frac{1}{2}\,(1-\eta_k^2)\;h_k\;\|F(x^k)\|\;. \tag{1.115}$$

The above inequality (1.115) implies the **a posteriori estimate**

$$[h_k]_1 \;:=\; \frac{2\,\|F(x^{k+1}) - r^k\|}{(1-\eta_k^2)\|F(x^k)\|} \;\leq\; h_k \tag{1.116}$$

and the **a priori estimate**

$$[h_{k+1}] \;:=\; \Theta_k\,[h_k]_1 \;\leq\; h_{k+1}\;. \tag{1.117}$$

Based on (1.117) we define

$$\overline{\eta}_{k+1} \;:=\; \overline{\Theta} \;-\; \frac{1}{2}\,[h_{k+1}]\;. \tag{1.118}$$

If we find

$$\overline{\eta}_{k+1} \;<\; \eta_k \tag{1.119}$$

with η_k from (1.113), we **continue the iteration in the quadratic convergence mode**.
Otherwise, we realize the **linear convergence mode** with some

$$\eta_{k+1} \;\leq\; \overline{\eta}_{k+1}\;. \tag{1.120}$$

3.3 Affine Conjugate Inexact Newton Methods

3.3.1 PCG (Preconditioned Conjugate Gradient)

The **Preconditioned Conjugate Gradient Method (PCG)** is an iterative solver for linear algebraic systems with a symmetric positive definite coefficient matrix $A \in \mathbb{R}^{n\times n}$. We recall that any symmetric positive definite matrix $C \in \mathbb{R}^{n\times n}$ defines an energy inner product $(\cdot,\cdot)_C$ according to

$$(u,v)_C \;:=\; (u,Cv) \quad , \quad u,v \in \mathbb{R}^n \; .$$

The associated energy norm is denoted by $\|\cdot\|_C$.
The PCG Method with a **symmetric positive definite preconditioner** $B \in \mathbb{R}^{n\times n}$ corresponds to the CG Method applied to the transformed linear algebraic system

$$B^{1/2}AB^{1/2}(B^{-1/2}y) \;=\; B^{1/2}b \; .$$

The **PCG Method** is implemented as follows:

PCG Initialization:

Given an **initial guess** $y_0 \in \mathbb{R}^n$, compute the **residual** $r_0 = b - Ay_0$ and the **preconditioned residual** $\overline{r}_0 = Br_0$ and set

$$p_0 \;:=\; \overline{r}_0 \quad , \quad \sigma_0 \;:=\; (r_0,\overline{r}_0) \;=\; \|r_0\|_B^2 \; .$$

PCG Iteration Loop: For $0 \le i \le i_{max}$ compute:

$$y_{i+1} \;=\; y_i \;+\; \frac{1}{\alpha_i}\, p_i \; ,$$

$$r_{i+1} \;=\; r_i \;-\; \frac{1}{\alpha_i}\, Ap_i \quad , \quad \overline{r}_{i+1} \;=\; Br_{i+1} \quad , \quad \alpha_i \;=\; \frac{\|p_i\|_A^2}{\sigma_i}$$

$$\gamma_i^2 \;=\; \frac{\sigma_i}{\alpha_i} \quad (=\; \|y_{i+1}-y_i\|_A^2) \; ,$$

$$p_{i+1} \;=\; \overline{r}_{i+1} \;+\; \frac{\sigma_{i+1}}{\sigma_i}\, p_i \quad , \quad \sigma_{i+1} \;=\; \|r_{i+1}\|_B^2 \; .$$

PCG minimizes the **energy error norm**

$$\|y - y_i\|_A = \min_{z \in \mathcal{K}_i(r_0, A)} \|y - z\|_A \,, \tag{1.121}$$

where $\mathcal{K}_i(r_0, A)$ denotes the **Krylov subspace**

$$\mathcal{K}_i(r_0, A) := \text{span}\{r_0, ..., A^{i-1} r_0\} \,. \tag{1.122}$$

PCG satisfies the **Galerkin orthogonality**

$$(y_i - y_0, y_{i+m} - y_i)_A = 0 \quad , \quad m \in \mathbb{N} \,. \tag{1.123}$$

Denoting by $y^* \in \mathbb{R}^n$ the unique solution of $Ay = b$ and by $\varepsilon_i := \|y^* - y_i\|_A^2$ the square of the iteration error in the energy norm, we have the following error representation:

Lemma 3.2 Representation of the iteration error

The PCG iteration error satisfies

$$\varepsilon_i = \sum_{j=i}^{n-1} \gamma_j^2 \,. \tag{1.124}$$

Proof. For $m = 1$ the Galerkin orthogonality implies the **orthogonal decompositions**

$$\|y_{i+1} - y_0\|_A^2 = \underbrace{\|y_{i+1} - y_i\|_A^2}_{= \gamma_i^2} + \|y_i - y_0\|_A^2 \,, \tag{1.125}$$

$$\|y_i - y_0\|_A^2 = \sum_{j=0}^{i-1} \|y_{j+1} - y_j\|_A^2 = \sum_{j=0}^{i-1} \gamma_j^2 \,. \tag{1.126}$$

On the other hand, observing $y_n = y^*$, for $m = n - i$ the Galerkin orthogonality yields

$$\underbrace{\|y^* - y_0\|_A^2}_{= \sum_{j=0}^{n-1} \gamma_j^2} = \underbrace{\|y^* - y_i\|_A^2}_{= \varepsilon_i^2} + \underbrace{\|y_i - y_0\|_A^2}_{= \sum_{j=0}^{i-1} \gamma_j^2} \,. \tag{1.127}$$

•

Computable lower bound for the iteration error

A lower bound for the iteration error in the energy norm is obviously given by

$$[\varepsilon_i] \;=\; \sum_{j=0}^{i+m} \gamma_j^2 \;. \tag{1.128}$$

In the **inexact Newton PCG method** we will control the inner PCG iterations by the **relative energy error norms**

$$\delta_i \;=\; \frac{\|y^* - y_i\|_A}{\|y_i\|_A} \;\approx\; \frac{\sqrt{[\varepsilon_i]}}{\|y_i\|_A} \tag{1.129}$$

and use the **termination criterion**

$$\delta_i \;\leq\; \overline{\delta} \;, \tag{1.130}$$

where $\overline{\delta}$ is a user specified accuracy.

3.3.2 Convergence of affine conjugate inexact Newton methods

We denote by $\delta x^k \in \mathbb{R}^n$ the result of the inner PCG iteration. As **initial value** for PCG we choose

$$\delta x_0^k \;=\; 0 \;. \tag{1.131}$$

Again, we will drop the subindices i for the inner PCG iterations and refer to η_k as the final value of the inner iterations at each outer iteration step k. We recall the **Galerkin orthogonality** (cf. (1.123))

$$(\delta x^k, F'(x^k)(\delta x^k - \Delta x^k)) \;=\; (\delta x^k, r^k) \;=\; 0 \;. \tag{1.132}$$

Theorem 3.5 Affine conjugate convergence theorem for the inexact Newton PCG method. Part I: Linear convergence

Suppose that $f : D \subset \mathbb{R}^n \to \mathbb{R}$ is a twice continuously differentiable strictly convex functional on D with the first derivative $F := f'$ and the Hessian $F' = f''$ which is symmetric and uniformly positive definite. Assume that $x^0 \in D$ is some initial guess such that the level set

$$\mathcal{L}_0 \;:=\; \{x \in D \mid f(x) \;\leq\; f(x^0)\}$$

is compact.
Let further the following **affine conjugate Lipschitz condition** be satisfied

$$\begin{aligned} &\|F'(z)^{-1/2}\Big(F'(y)-F'(x)\Big)v\| \;\leq \\ &\leq\; \omega\,\|F'(x)^{1/2}(y-x)\|\,\|F'(x)^{1/2}v\|\;,\;\; x,y,z\in D\;,\;\omega\geq 0\;. \end{aligned} \tag{1.133}$$

For the inner Newton PCG iterations consider the exact error terms

$$\varepsilon_k \;:=\; \|F'(x^k)^{1/2}\Delta x^k\|^2$$

and the **Kantorovich quantities**

$$h_k \;:=\; \omega\,\|F'(x^k)^{1/2}\Delta x^k\|$$

as well as their inexact analogues

$$\varepsilon_k^\delta \;:=\; \|F'(x^k)^{1/2}\delta x^k\|^2 \;=\; \frac{\varepsilon_k}{1+\delta_k^2}$$

and

$$h_k^\delta \;:=\; \omega\,\|F'(x^k)^{1/2}\delta x^k\| \;=\; \frac{h_k}{\sqrt{1+\delta_k^2}}\;,$$

where δ_k characterizes the **inner PCG iteration error**

$$\delta_k \;:=\; \frac{\|F'(x^k)^{1/2}\Big(\delta x^k-\Delta x^k\Big)\|}{\|F'(x^k)^{1/2}\delta x^k\|}\;.$$

Assume that for some $\overline{\Theta}<1$

$$h_0 \;<\; 2\,\overline{\Theta} \;<\; 2 \tag{1.134}$$

and that

$$\delta_{k+1} \;\geq\; \delta_k \quad,\quad k\in\mathbb{N}_0 \tag{1.135}$$

holds true throughout the outer Newton iterations.
Control the inner iterations according to

$$\vartheta(h_k^\delta,\delta_k) \;:=\; \frac{h_k^\delta+\delta_k\Big(h_k^\delta+\sqrt{4+(h_k^\delta)^2}\Big)}{2\sqrt{1+\delta_k^2}} \;\leq\; \overline{\theta}\;. \tag{1.136}$$

Then, the **inexact Newton PCG iterates** $x^k, k \in \mathbb{N}_0$ stay in $\mathcal{L}_0$ and **converge linearly** to some $x^* \in \mathcal{L}_0$ with $f(x^*) = \min_{x \in D} f(x)$.
The following estimates hold true

$$\|F'(x^{k+1})^{1/2} \Delta x^{k+1}\| \;\leq\; \overline{\Theta}\, \|F'(x^k)^{1/2} \Delta x^k\| \quad , \quad k \in \mathbb{N}_0 \;, \tag{1.137}$$

$$\|F'(x^{k+1})^{1/2} \delta x^{k+1}\| \;\leq\; \overline{\Theta}\, \|F'(x^k)^{1/2} \delta x^k\| \quad , \quad k \in \mathbb{N}_0 \;. \tag{1.138}$$

Moreover, the **objective functional** is reduced according to

$$-\,\frac{1}{10}\, h_k^\delta\, \varepsilon_k^\delta \;\leq\; f(x^k) \;-\; f(x^{k+1}) \;-\; \frac{2}{3}\, \varepsilon_k^\delta \;\leq\; \frac{1}{10}\, h_k^\delta\, \varepsilon_k^\delta \;. \tag{1.139}$$

Proof. Observing

$$r^k \;=\; F(x^k) \;+\; F'(x^k) \delta x^k \quad , \quad k \in \mathbb{N}_0 \;,$$

for $\lambda \in [0,1]$ we obtain

$$f(x^k + \lambda \delta x^k) \;-\; f(x^k) \;=\; \lambda \int_{s=0}^{\lambda} (\delta x^k, F(x^k + s \delta x^k))\; ds \;= \tag{1.140}$$

$$= \;\lambda \int_{s=0}^{\lambda} (\delta x^k, F(x^k + s \delta x^k) \;- F(x^k))\; ds \;+\; \lambda \int_{s=0}^{\lambda} (\delta x^k, F(x^k))\; ds \;=$$

$$= \;\lambda \int_{s=0}^{\lambda} s \int_{t=0}^{s} (\delta x^k, F'(x^k + st \delta x^k) \delta x^k)\; dt\; ds \;+\; \lambda \int_{s=0}^{\lambda} (\delta x^k, F(x^k))\; ds \;=$$

$$= \;\lambda \int_{s=0}^{\lambda} s \int_{t=0}^{s} (\delta x^k, \Big(F'(x^k + st \delta x^k) - F'(x^k)\Big) \delta x^k)\; dt\; ds \;+$$

$$+\; \lambda \int_{s=0}^{\lambda} s \int_{t=0}^{s} (\delta x^k, F'(x^k) \delta x^k)\; dt\; ds \;+\; \lambda \int_{s=0}^{\lambda} (\delta x^k, \underbrace{F(x^k)}_{r^k - F'(x^k)\delta x^k})\; ds \;=$$

$$= \lambda \int_{s=0}^{\lambda} s \int_{t=0}^{s} \underbrace{(F'(x^k)^{1/2}\delta x^k, F'(x^k)^{-1/2}\Big(F'(x^k + st\delta x^k) - F'(x^k)\Big)\delta x^k)}_{\leq \|F'(x^k)^{1/2}\delta x^k\| \ \omega \ s \ t \ \|F'(x^k)^{1/2}\delta x^k\|^2 \ = \ s \ t \ h_k^\delta \ \varepsilon_k^\delta} \ dt \ ds$$

$$+ \lambda \int_{s=0}^{\lambda} s \int_{t=0}^{s} (\delta x^k, F'(x^k)\delta x^k) \ dt \ ds \ - \ \lambda \int_{s=0}^{\lambda} (\delta x^k, F'(x^k)\delta x^k) \ ds \ +$$

$$+ \lambda \int_{s=0}^{\lambda} \underbrace{(\delta x^k, r^k)}_{= \ 0 \ due \ to \ (1.123)} \ ds \ \leq \ \frac{1}{10} \ \lambda^6 \ h_k^\delta \ \varepsilon_k^\delta \ + \ \frac{1}{3} \ \lambda^4 \ \varepsilon_k^\delta \ - \ \lambda^2 \ \varepsilon_k^\delta \ .$$

It readily follows from (1.140) that

$$f(x^k + \lambda\delta x^k) \ \leq \ f(x^k) \ + \ \lambda^2 \ (\frac{1}{10} \ h_k^\delta \ \varepsilon_k^\delta \ + \ (\frac{1}{3} \ \lambda^2 \ - \ 1) \ \varepsilon_k^\delta) \ . \qquad (1.141)$$

Denoting by $\mathcal{L}_k$ the level set

$$\mathcal{L}_k \ := \ \{ \ x \in D \mid f(x) \ \leq \ f(x^k) \ \} \ ,$$

by **induction on** k we prove

$$h_k \ < \ 2 \quad \text{and hence,} \quad x^{k+1} \in \mathcal{L}_k \ . \qquad (1.142)$$

For $k = 0$, we have $h_0 < 2$ by assumption (1.134). Since $h_0^\delta \leq h_0$, (1.141) readily shows $f(x^1) < f(x^0)$, whence $x^1 \in \mathcal{L}_0$.
Now, assuming (1.142) to hold true for some $k \in \mathbb{N}$, again taking advantage of $h_k^\delta \leq h_k < 2$, (1.141) yields $f(x^{k+1}) < f(x^k)$ and thus $x^{k+1} \in \mathcal{L}_k$.
Moreover, choosing $\lambda = 1$ in (1.141), we obtain the left-hand side of the **functional descent property** (1.139). We note that we get the right-hand side of (1.139), if in (1.140) we estimate by the other direction of the Cauchy-Schwarz inequality.
Finally, in order to prove the **contraction properties** (1.137),(1.138) and linear convergence, we estimate the local energy norms as follows:

$$\|F'(x^{k+1})^{1/2}\Delta x^{k+1}\| \ = \ \|F'(x^{k+1})^{-1/2} \ \underbrace{F'(x^{k+1})\Delta x^{k+1}}_{= \ -F(x^{k+1})}\| \ =$$

$$= \ \|F'(x^{k+1})^{-1/2} \ \Big(F(x^{k+1}) \ \pm \ F(x^k)\Big)\| \ =$$

$$= \|F'(x^{k+1})^{-1/2}\left(F(x^{k+1}) - F(x^k)\right) + F'(x^{k+1})^{-1/2}\,F(x^k)\| \,.$$

Observing

$$F(x^k) = -F'(x^k)\delta x^k + r^k \,,$$

and using the **affine conjugate Lipschitz condition** we obtain

$$\|F'(x^{k+1})^{1/2}\Delta x^{k+1}\| = \tag{1.143}$$

$$= \|F'(x^{k+1})^{-1/2}\left(\int_0^1 \left(F'(x^k + t\delta x^k) - F'(x^k)\right)\delta x^k\, dt + r^k\right)\| \leq$$

$$\leq \frac{1}{2}\,\omega\,\|F'(x^k)^{1/2}\delta x^k\|^2 + \|F'(x^{k+1})^{-1/2}r^k\| \,.$$

Setting $z = \delta x^k - \Delta x^k$, for the second term on the right-hand side of the previous inequality we get the implicit estimate

$$\|F'(x^{k+1})^{-1/2}r^k\|^2 \leq$$

$$\leq \|F'(x^k)^{1/2}z\|^2 + h_k^\delta\,\|F'(x^k)^{1/2}z\|\,\|F'(x^{k+1})^{-1/2}r^k\| \,,$$

which gives the explicit bound

$$\|F'(x^{k+1})^{-1/2}r^k\| \leq \frac{1}{2}\left(h_k^\delta + \sqrt{4 + (h_k^\delta)^2}\right)\|F'(x^k)z\| \,. \tag{1.144}$$

Using (1.144) in (1.143) results in

$$\omega\,\|F'(x^{k+1})^{1/2}\Delta x^{k+1}\| \leq$$

$$\leq \frac{1}{2}\underbrace{\omega^2\,\|F'(x^k)^{1/2}\delta x^k\|^2}_{=\,(h_k^\delta)^2} + \frac{1}{2}\left(h_k^\delta + \sqrt{4 + (h_k^\delta)^2}\right)\underbrace{\omega\,\|F'(x^k)^{1/2}z\|}_{=\,\delta_k\;h_k^\delta} \,.$$

Taking (1.136) into account, we thus get the **contraction factor estimate**

$$\Theta_k := \frac{\omega\,\|F'(x^{k+1})^{1/2}\Delta x^{k+1}\|}{\underbrace{\omega\,\|F'(x^k)^{1/2}\Delta x^k\|}_{=\,h_k\,=\,\sqrt{1+\delta_k^2}\;h_k^\delta}} \leq \vartheta(h_k^\delta, \delta_k) \leq \overline{\Theta} \,, \tag{1.145}$$

which proves (1.137) and **linear convergence**.
For the proof of (1.138) we observe

$$\|F'(x^\ell)^{1/2}\Delta x^\ell\|^2 \;=\; (1+\delta_\ell^2)\;\|F'(x^\ell)^{1/2}\delta x^\ell\|^2 \quad , \quad \ell = k, k+1 \;,$$

as well as $\delta_{k+1} \geq \delta_k$ and obtain

$$\frac{\|F'(x^{k+1})^{1/2}\delta x^{k+1}\|}{\|F'(x^k)^{1/2}\delta x^k\|} \;\leq\; \sqrt{\frac{1+\delta_k^2}{1+\delta_{k+1}^2}}\;\Theta_k \;\leq\; \Theta_k \;\leq\; \overline{\Theta}\;. \tag{1.146}$$

By standard arguments we further show that the sequence $\{x^k\}_{\mathbb{N}_0}$ of inexact Newton PCG iterates is a Cauchy sequence in $\mathcal{L}_0$ and there exists an $x^* \in \mathcal{L}_0$ such that $x^k \to x^*$ $(k \to \infty)$ with $F(x^*) = 0$.

•

Theorem 3.6 Affine conjugate convergence theorem for the inexact Newton PCG method. Part II: Quadratic convergence

Under the same assumptions on $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$ as in Theorem 3.5 suppose that the initial guess $x^0 \in D$ satisfies

$$h_0^\delta \;<\; \frac{2}{1+\rho} \tag{1.147}$$

for some appropriate $\rho > 0$ and **control the inner iterations** such that

$$\delta_k \;\leq\; \frac{\rho}{2}\;\frac{h_k^\delta}{h_k^\delta + \sqrt{4+(h_k^\delta)^2}}\;. \tag{1.148}$$

Then, there holds:

(i) The **Newton CGNE iterates** $x^k, k \in \mathbb{N}_0$ stay in $\mathcal{L}_0$ and **converge quadratically** to some $x^* \in \mathcal{L}_0$ with $F(x^*) = 0$.

(ii) The **exact Newton increments** and the **inexact Newton increments** decrease quadratically according to

$$\|F'(x^{k+1})^{1/2}\;\Delta x^{k+1}\| \;\leq\; \frac{1+\rho}{2}\;\omega\;\|F'(x^k)^{1/2}\;\Delta x^k\|^2\;, \tag{1.149}$$

$$\|F'(x^{k+1})^{1/2}\;\delta x^{k+1}\| \;\leq\; \frac{1+\rho}{2}\;\omega\;\|F'(x^k)^{1/2}\;\delta x^k\|^2\;. \tag{1.150}$$

Proof. Using (1.148) in (1.145) yields

$$\frac{\|F'(x^{k+1})^{1/2}\Delta x^{k+1}\|}{\|F'(x^k)^{1/2}\Delta x^k\|} \;\leq\; \frac{h_k^\delta \;+\; \delta_k\,(h_k^\delta + \sqrt{1+(h_k^\delta)^2})}{2\,\sqrt{1+\delta_k^2}} \;\leq\; \frac{1}{2}\,(1+\rho)\,h_k^\delta\,,$$

which proves (1.149) in view of $h_k^\delta \leq h_k \leq h_0 < 2\overline{\Theta}$.
The proof of (1.150) follows along the same line by using (1.148) in (1.146).
•

3.3.3 Algorithmic aspects of the affine conjugate inexact Newton PCG method

(i) Convergence monitor

Let us assume that the quantity $\overline{\Theta} < 1$ in both the **linear convergence mode** and the **quadratic convergence mode** has been specified and let us further assume that we use the startiterate $\delta x_0^k = 0$ in the inner PCG iteration.
Denoting by $\overline{\delta}_k$ an easily computable estimate of the **relative energy norm iteration error** δ_k, we accept a new iterate x^{k+1}, if the condition

$$f(x^{k+1}) \;-\; f(x^k) \;\leq\; -\,\frac{1}{10}\,\varepsilon_k \;=\; -\,\frac{1}{10}\,(1+\overline{\delta}_k^2)\varepsilon_k^\delta \tag{1.151}$$

or the **monotonicity test**

$$\Theta_k \;:=\; \Big(\frac{\varepsilon_{k+1}}{\varepsilon_k}\Big)^{1/2} \;=\; \Big(\frac{(1+\overline{\delta}_{k+1}^2)\;\varepsilon_{k+1}^\delta}{(1+\overline{\delta}_k^2)\;\varepsilon_k^\delta}\Big)^{1/2} \;\leq\; \overline{\Theta} \;<\; 1 \tag{1.152}$$

is satisfied. We consider the outer iteration as **divergent**, if neither (1.151) nor (1.152) hold true.

(ii) Termination criterion

With respect to a user specified accuracy ETOL, the inexact Newton PCG iteration will be terminated, if either

$$\varepsilon_k \;=\; (1 \;+\; \overline{\delta}_k^2)\;\varepsilon_k^\delta \;\leq\; \mathrm{ETOL}^2\,. \tag{1.153}$$

or

$$f(x^k) \;-\; f(x^{k+1}) \;\leq\; \frac{1}{2}\,\mathrm{ETOL}^2\,. \tag{1.154}$$

(iii) Balancing outer and inner iterations

For $k = 0$, we choose $\overline{\delta}_0 = \delta_0 = \frac{1}{4}$.
As in case of the inexact Newton CGNE iteration, for $k \geq 1$ we begin with the

quadratic convergence mode and switch to the linear convergence mode as soon as the approximate contraction factor $\tilde{\Theta}_k$ is below some prespecified threshold value $\overline{\Theta} \leq \frac{1}{2}$.

(iii)$_1$ Quadratic convergence mode

A computationally realizable **termination criterion for the inner PCG iteration in the quadratic convergence mode** is given by

$$\overline{\delta}_k \;\leq\; \frac{\rho\,[h_k^\delta]}{[h_k^\delta] \;+\; \sqrt{4+[h_k^\delta]^2}}\;, \tag{1.155}$$

where $[h_k^\delta]$ is an appropriate **a priori estimate** of the inexact Kantorovich quantity h_k^δ. In view of (1.145), we have the **a posteriori estimates**

$$[h_k^\delta]_2 \;:=\; \frac{10}{\varepsilon_k^\delta}\,|f(x^{k+1}) \;-\; f(x^k) \;+\; \frac{1}{3}\,\varepsilon_k^\delta| \tag{1.156}$$

and

$$[h_k]_2 \;:=\; \sqrt{1+\overline{\delta}_k^2}\;|[h_k^\delta]_2\;. \tag{1.157}$$

We note that (1.157) yields the **a priori estimate**

$$[h_k] \;:=\; \Theta_{k-1}\;[h_{k-1}]_2\;. \tag{1.158}$$

Using (1.158) in (1.157), for the inexact Kantorovich quantity we obtain the following **a priori estimate**

$$[h_k^\delta] \;:=\; \frac{[h_k]}{\sqrt{1+\overline{\delta}_k^2}}\;. \tag{1.159}$$

Inserting (1.159) into (1.155), we obtain a simple nonlinear equation in $\overline{\delta}_k$.

Remark 3.5 Computational work in the quadratic convergence mode

Since $\overline{\delta}_k \to 0$ $(k \to \infty)$ is enforced, it follows that:
The more the iterates x^k approach the solution x^*, the more computational work is required for the inner iterations to guarantee quadratic convergence of the outer iteration.

(iii)$_2$ Linear convergence mode

We switch to the linear convergence mode, if

$$\tilde{\Theta}_k \;<\; \overline{\Theta} \tag{1.160}$$

is satisfied.
The computationally realizable **termination criterion for the inner iteration in the linear convergence mode** is

$$[\vartheta(h_k^\delta, \delta_k)] \;:=\; \vartheta([h_k^\delta], \overline{\delta}_k) \;\leq\; \overline{\Theta}\;. \tag{1.161}$$

Since asymptotically there holds

$$\overline{\delta}_k \to \frac{\overline{\Theta}}{\sqrt{1-\overline{\Theta}^2}} \quad (k \to \infty)\;,$$

we observe:

Remark 3.6 Computational work in the linear convergence mode
The more the iterates x^k approach the solution x^*, the less computational work is required for the inner iterations to guarantee linear convergence of the outer iteration.

4. Quasi-Newton Methods

4.1 Introduction

Given $F : D \subset \mathbb{R}^n \times \mathbb{R}^n$ as well as $x^k, x^{k+1} \in D$, $x^k \neq x^{k+1}$, the idea is to **approximate F locally** around x^{k+1} by an **affine function**

$$S_{k+1}(x) \;:=\; F(x^{k+1}) \;+\; J_{k+1}(x - x^{k+1}) \;,\; J_{k+1} \in \mathbb{R}^{n\times n} \;, \tag{1.162}$$

such that

$$S_{k+1}(x^k) \;=\; F(x^k) \;. \tag{1.163}$$

The requirement (1.163) gives rise to the so-called **secant condition**

$$J \underbrace{\left(x^{k+1} - x^k\right)}_{=:\,\delta x^k} \;=\; \underbrace{F(x^{k+1}) - F(x^k)}_{=:\,y^k} \;. \tag{1.164}$$

The matrix J is not uniquely determined by (1.164), since

$$\dim \mathcal{S}_{k+1} \;=\; (n-1)n \;, \tag{1.165}$$

where

$$\mathcal{S}_{k+1} \;:=\; \{J \in \mathbb{R}^{n\times n} \mid J\delta x^k \;=\; y^k\} \;. \tag{1.166}$$

There are different criteria to select an appropriate $J \in \mathcal{S}_{k+1}$.

4.1.1 The Good Broyden rank 1 update

Let us consider the **change in the affine model** as given by

$$S_{k+1}(x) - S_k(x) \;=\; (J_{k+1} - J_k)(x - x^k) \;. \tag{1.167}$$

An appropriate idea is to choose $J_{k+1} \in \mathcal{S}_{k+1}$ such that there is a **least change in the affine model** in the sense

$$\|J_{k+1} - J_k\|_F \;=\; \min_{J\in\mathcal{S}_{k+1}} \|J - J_k\|_F \;, \tag{1.168}$$

where $\|\cdot\|_F$ stands for the **Frobenius norm** (observe $J = (J_{ik})_{i,k=1}^n$)

$$\|J\|_F \;:=\; \Big(\sum_{i,k=1}^{n} J_{ik}^2\Big)^{1/2} \;. \tag{1.169}$$

The solution of (1.169) can be heuristically motivated as follows: Choose $t^k \perp \delta x^k$ such that

$$x - x^k \;=\; \alpha \delta x^k \;+\; t^k \;.$$

Then, (1.167) reads

$$S_{k+1}(x) - S_k(x) \;=\; \underbrace{\alpha (J_{k+1} - J_k)\delta x^k}_{= \alpha(y^k - J_k \delta x^k)} + (J_{k+1} - J_k) t^k \;. \tag{1.170}$$

Now, choose $J_{k+1} \in \mathcal{S}_{k+1}$ such that

$$(J_{k+1} - J_k) t^k \;=\; 0 \;.$$

It follows that

$$\text{rank } (J_{k+1} - J_k) \;=\; 1 \quad , \quad J_{k+1} - J_k \;=\; v^k (\delta x^k)^T \;. \tag{1.171}$$

Inserting (1.171) into (1.170) yields

$$\alpha \; v^k \; (\delta x^k)^T \delta x^k \;=\; \alpha \; (y^k - J_k \delta x^k) \;,$$

which results in

$$v^k \;=\; \frac{y^k - J_k \delta x^k}{(\delta x^k)^T \delta x^k} \;.$$

Altogether, this gives us **Broyden's rank 1 update (Good Broyden)**

$$J_{k+1} \;=\; J_k \;+\; \left[\; F(x^{k+1}) - F(x^k) - J_k \delta x^k \; \right] \frac{(\delta x^k)^T}{(\delta x^k)^T \delta x^k} \;. \tag{1.172}$$

For the solution of nonlinear systems, we are more interested in **updates of the inverse** of J_k. Such an update can be provided by the **Sherman-Morrison-Woodbury formula**

$$(A \;+\; uv^T)^{-1} \;=\; A^{-1} \;-\; \frac{A^{-1} u v^T A^{-1}}{1 + v^T A^{-1} u} \;. \tag{1.173}$$

Setting

$$A \;:=\; J_k \quad , \quad u \;:=\; F(x^{k+1}) - F(x^k) - J_k \delta x^k \quad , \quad v \;:=\; \frac{(\delta x^k)^T}{(\delta x^k)^T \delta x^k} \;,$$

we obtain

$$J_{k+1}^{-1} \;=\; J_k^{-1} \;+\; \frac{\left[\delta x^k - J_k^{-1} (F(x^{k+1}) - F(x^k)) \right] (\delta x^k)^T J_k^{-1}}{(\delta x^k)^T J_k^{-1} \left[F(x^{k+1}) - F(x^k) \right]} \;. \tag{1.174}$$

4.1.2 The Bad Broyden rank 1 update

Instead of (1.168), an alternative to choose $J_{k+1} \in \mathcal{S}_{k+1}$ such that there is a **least change in the solution of the affine model**, i.e.,

$$\|J_{k+1}^{-1} - J_k^{-1}\|_F = \min_{J \in \mathcal{S}_{k+1}} \|J^{-1} - J_k^{-1}\|_F \,. \tag{1.175}$$

Similar considerations as before lead us to the **Broyden's alternative rank 1 update (Bad Broyden)**

$$J_{k+1}^{-1} = J_k^{-1} + \frac{\left[\delta x^k - J_k^{-1}\Big(F(x^{k+1}) - F(x^k)\Big)\right]\Big(F(x^{k+1}) - F(x^k)\Big)^T}{\Big(F(x^{k+1}) - F(x^k)\Big)^T\Big(F(x^{k+1}) - F(x^k)\Big)} \,. \tag{1.176}$$

4.2 Affine covariant Quasi-Newton method

4.2.1 Affine covariant Quasi-Newton convergence theory

Affine covariant Quasi-Newton methods require the secant condition (1.164) to be stated by means of affine covariant terms in the domain of definition of the nonlinear mapping F.
Observing that we compute the Quasi-Newton increment δx^k as the solution of

$$J_k \delta x^k = -F(x^k) \,, \tag{1.177}$$

we can rewrite (1.164) according to

$$(J_k - J)\delta x^k = -F(x^{k+1}) \,.$$

Multiplication by J_k^{-1} yields the **affine covariant secant condition**

$$\overline{\delta} x^{k+1} := \underbrace{(I - J_k^{-1}J)}_{=:\, E_k(J)} \delta x^k = -J_k^{-1} F(x^{k+1}) \,. \tag{1.178}$$

we note that any **rank 1 update** of the form

$$\tilde{J}_{k+1} = J_k \left(I - \frac{\overline{\delta} x^{k+1} v^T}{v^T \delta x^k}\right) \quad , \quad v \in \mathbb{R}^n \setminus \{0\} \tag{1.179}$$

satisfies the affine covariant secant condition (1.178).
In particular, for $v = \delta x^k$ we recover the **Good Broyden**.

Theorem 4.1 Properties of the affine covariant Quasi-Newton method

For Broyden's affine covariant rank 1 update (Good Broyden)

$$J_{k+1} \;=\; J_k \left(I \;-\; \frac{\overline{\delta} x^{k+1} (\delta x^k)^T}{\|\delta x^k\|^2} \right) \tag{1.180}$$

assume that the **local contraction condition**

$$\Theta_k \;=\; \frac{\|\overline{\delta} x^{k+1}\|}{\|\delta x^k\|} \;<\; \frac{1}{2} \tag{1.181}$$

is satisfied. Then, there holds:

(i) The update matrix J_{k+1} is a **least change update** in the sense that

$$\|E_k(J_{k+1})\| \;\leq\; \|E_k(J)\| \quad , \quad J \in \mathcal{S}_{k+1} \;, \tag{1.182}$$

$$\|E_k(J_{k+1})\| \;\leq\; \Theta_k \;. \tag{1.183}$$

(ii) If J_k is regular, then J_{k+1} is regular as well with the **inverse** given by

$$J_{k+1}^{-1} \;=\; \left(I \;+\; \frac{\overline{\delta} x^{k+1} (\delta x^k)^T}{(1-\alpha_{k+1})\, \|\delta x^k\|^2} \right) J_k^{-1} \;, \tag{1.184}$$

where

$$\alpha_{k+1} \;=\; \frac{(\delta x^k)^T \overline{\delta} x^{k+1}}{\|\delta x^k\|^2} \;<\; \frac{1}{2} \;.$$

(iii) The **Quasi-Newton increment** δx^{k+1} is given by

$$\delta x^{k+1} \;=\; -\; J_{k+1}^{-1} F(x^{k+1}) \;=\; \frac{\overline{\delta} x^{k+1}}{1-\alpha_{k+1}} \;. \tag{1.185}$$

(iv) The **Quasi-Newton increments decrease** according to

$$\frac{\|\delta x^{k+1}\|}{\|\delta x^k\|} \;\leq\; \frac{\Theta_k}{1-\alpha_{k+1}} \;<\; 1 \;. \tag{1.186}$$

Proof. In view of (1.178) we have

$$\|E_k(J_{k+1})\| \;=\; \|\frac{\overline{\delta} x^{k+1} (\delta x^k)^T}{\|\delta x^k\|^2}\| \;=\; \|E_k(J)\, \frac{\delta x^k (\delta x^k)^T}{\|\delta x^k\|^2}\| \;\leq\; \|E_k(J)\| \;,$$

which proves (1.182). Moreover, (1.183) follows readily from

$$\|E_k(J_{k+1})\| \;=\; \|\frac{\overline{\delta}x^{k+1}(\delta x^k)^T}{\|\delta x^k\|^2}\| \;\leq\; \frac{\|\overline{\delta}x^{k+1}\|}{\|\delta x^k\|} \;=\; \Theta_k \;.$$

The same argument shows

$$|\alpha_{k+1}| \;\leq\; \Theta_k \;<\; \frac{1}{2}\;,$$

and hence, (1.186) follows from

$$\frac{\|\delta x^{k+1}\|}{\|\delta x^k\|} \;=\; \frac{\Theta_k}{1-\alpha_{k+1}} \;\leq\; \frac{\Theta_k}{1-\Theta_k} \;<\; 1\;.$$

Finally, the proofs of **(ii)** and **(iii)** are direct consequences of the Sherman-Morrison-Woodbury formula (1.173).

•

Theorem 4.2 Convergence of the affine covariant Quasi-Newton method

Suppose that that $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$, $D \subset \mathbb{R}^n$ convex, is continuously differentiable on D. Let $x^* \in D$ be the unique solution of $F(x) = 0$ in D with invertible Jacobian $F'(x^*)$. Assume that the following **affine covariant Lipschitz condition** is satisfied

$$\|F'(x^*)^{-1}\Big(F'(x) - F'(x^*)\Big)v\| \;\leq\; \omega\;\|x - x^*\|\;\|v\|\;, \tag{1.187}$$

where $x, x+v \in D, v \in \mathbb{R}^n$.
For some $0 < \overline{\Theta} < 1$ assume further that:

(a) The **initial approximate Jacobian** J_0 satisfies

$$\delta_0 \;:=\; \|F'(x^*)^{-1}\;(J_0 \;-\; F'(x^0))\| \;<\; \frac{\overline{\Theta}}{1+\overline{\Theta}}\;. \tag{1.188}$$

(b) The **initial guess** $x^0 \in D$ satisfies

$$t_0 \;:=\; \omega\;\|x^0 - x^+\| \;\leq\; \frac{1-\overline{\Theta}}{2-\overline{\Theta}}\;\Big(\frac{\overline{\Theta}}{1+\overline{\Theta}} \;-\; \delta_0\Big)\;. \tag{1.189}$$

Then, there holds:

(i) The **Quasi-Newton iterates** $x^k, k \in \mathbb{N}_0$ converge to x^* according to

$$\|x^{k+1} \;-\; x^*\| \;<\; \overline{\Theta}\;\|x^k \;-\; x^*\|\;, \tag{1.190}$$

$$\|\delta x^{k+1}\| \;\leq\; \overline{\Theta}\;\|\delta x^k\|\;. \tag{1.191}$$

We have **superlinear convergence** in the sense that

$$\lim_{k\to\infty}\frac{\|\delta x^{k+1}\|}{\|\delta x^k\|} \;=\; 0\;. \tag{1.192}$$

(ii) For $E_k^* := F'(x^*)^{-1}J_k - I = F'(x^*)^{-1}(J_k - F'(x^*))$ the following **affine covariant bounded deterioration property** holds true

$$\|E_k^*\| \;\leq\; \frac{\overline{\Theta}}{1+\overline{\Theta}} \;<\; \frac{1}{2}\;. \tag{1.193}$$

Moreover, asymptotically we have

$$\lim_{k\to\infty}\frac{\|E_k^*\delta x^k\|}{\|\delta x^k\|} \;=\; 0\;. \tag{1.194}$$

Proof. Denoting by $e_k := x^k - x^*$ the **iteration error**, we set

$$t_k \;:=\; \omega\;\|e_k\|\;.$$

We first derive an estimate for the **contraction of the Quasi-Newton increments**. For this purpose, using the **affine covariant Lipschitz condition**, we get

$$\|F'(x^*)^{-1}F(x^{k+1})\| \;=\; \|F'(x^*)^{-1}\;\Big(F(x^{k+1})\pm F(x^k)\Big)\| \;= \tag{1.195}$$

$$=\; \|F'(x^*)^{-1}\;\Big(F(x^{k+1})-F(x^k)\Big) \;+\; \underbrace{F'(x^*)^{-1}\;F(x^k)}_{=\,-F'(x^*)^{-1}J_k\delta x^k\;=\;-E_k^*\delta x^k-\delta x^k}\|\;\leq$$

$$\leq\; \int_0^1 \|F'(x^*)^{-1}\;\Big(F'(x^k+t\delta x^k)\;-\;F'(x^*)\Big)\delta x^k\|\;dt\;+\;\|E_k^*\delta x^k\|\;\leq$$

$$\leq\; \omega\int_0^1 \|x^k\;+\;t\;\underbrace{\delta x^k}_{=\,x^{k+1}-x^k}\;-\;x^*\|\;dt\;\|\delta x^k\|\;+\;\|E_k^*\delta x^k\|\;\leq$$

$$\leq\; \omega\int_0^1 \Big(t\;\|x^k-x^*\|\;+\;(1-t)\;\|x^{k+1}-x^*\|\Big)\;dt\;\|\delta x^k\|\;+\;\|E_k^*\delta x^k\|\;\leq$$

$$\leq \left(\frac{1}{2}(t_{k+1}+t_k) \;+\; \frac{\|E_k^*\delta x^k\|}{\|\delta x^k\|}\right) \|\delta x^k\| \;.$$

On the other hand, assuming $\frac{\|E_{k+1}^*\delta x^{k+1}\|}{\|\delta x^{k+1}\|} < 1$ we obtain

$$\|F'(x^*)^{-1}F(x^{k+1})\| = \|(I+E_{k+1}^*)\delta x^{k+1}\| \geq (1-\frac{\|E_{k+1}^*\delta x^{k+1}\|}{\|\delta x^{k+1}\|})\;\|\delta x^{k+1}\| \quad (1.196)$$

Setting

$$\eta_\ell \;:=\; \frac{\|E_\ell^*\delta x^\ell\|}{\|\delta x^\ell\|}\;,\; \ell = k, k+1, \quad , \quad \bar{t}_k \;:=\; \frac{1}{2}\;(t_k+t_{k+1})\;,$$

the combination of (1.195) and (1.196) yields

$$\frac{\|\delta x^{k+1}\|}{\|\delta x^k\|} \;\leq\; \frac{\eta_k+\bar{t}_k}{1-\eta_{k+1}}\;. \quad (1.197)$$

Next, we establish an estimate for the **iterative error terms** t_k. We have

$$e_{k+1} \;=\; e_k \;+\; \delta x^k \;=\; e_k \;-\; J_k^{-1}F(x^k) \;=$$

$$=\; e_k \;-\; J_k^{-1}\left(F(x^k) \;-\; F(x^*)\right) \;=$$

$$=\; e_k \;-\; \underbrace{J_k^{-1}F'(x^*)}_{=\,(I+E_k^*)^{-1}}F'(x^*)^{-1}\left(F(x^k) \;-\; F(x^*)\right) \;=$$

$$=\; (I+E_k^*)^{-1}\left[(I+E_k^*)e_k \;-\; F'(x^*)^{-1}\left(F(x^k) \;-\; F(x^*)\right)\right] \;=$$

$$=\; (I+E_k^*)^{-1}\left[E_k^*e_k \;-\; F'(x^*)^{-1}\int_0^1 \left(F'(x^k+te_k)-F'(x^*)\right)e_k\;dt\right] .$$

Applying the affine covariant Lipschitz condition again, we arrive at

$$t_{k+1} \;\leq\; \frac{\|E_k^*\| \;+\; \frac{1}{2}t_k}{1 \;-\; \|E_k^*\|}\;t_k\;. \quad (1.198)$$

Comparing (1.197) and (1.198), we find

$$t_{k+1} \;<\; \overline{\Theta}_k\;t_k \quad , \quad \overline{\Theta}_k \;:=\; \frac{\|E_k^*\| \;+\; \bar{t}_k}{1 \;-\; \|E_k^*\|}. \quad (1.199)$$

As far as the **approximation properties of the rank 1 updates** are concerned, we start from

$$E_{k+1}^* \;=\; F'(x^*)^{-1} J_{k+1} \;-\; I \;= \tag{1.200}$$

$$= \; F'(x^*)^{-1} \; J_k \left(I \;-\; \frac{\overbrace{\delta x^{k+1}}^{= \, -J_k^{-1} F(x^{k+1})} (\delta x^k)^T}{\|\delta x^k\|^2} \right) \;-\; I \;=$$

$$= \; E_k^* \;+\; F'(x^*)^{-1} \; \frac{F(x^{k+1}) \; (\delta x^k)^T}{\|\delta x^k\|^2} \; .$$

We further have

$$F'(x^*)^{-1} \; F(x^{k+1}) \;=\; F'(x^*)^{-1} \left(F(x^{k+1}) \;-\; F(x^k) \right) \;+\; \underbrace{F'(x^*)^{-1} \; \underbrace{F(x^k)}_{= \, -J_k \delta x^k}}_{= \, -(I + E_k^*) \delta x^k} \;=$$

$$= \; F'(x^*)^{-1} \int_0^1 F'(x^k + t \delta x^k) \delta x^k \; dt \;-\; F'(x^*)^{-1} F'(x^*) \delta x^k \;-\; E_k^* \delta x^k \;=$$

$$= \; \underbrace{F'(x^*)^{-1} \int_0^1 \left(F'(x^k + t \delta x^k) \;-\; F'(x^*) \right) dt}_{=: \; D_{k+1}} \; \delta x^k \;-\; E_k^* \delta x^k \; .$$

Consequently, (1.200) yields

$$E_{k+1}^* \;=\; E_k^* \underbrace{\left(I \;-\; \frac{\delta x^k (\delta x^k)^T}{\|\delta x^k\|^2} \right)}_{=: \; Q_k} \;+\; D_{k+1} \; \underbrace{\frac{\delta x^k (\delta x^k)^T}{\|\delta x^k\|^2}}_{=: \; I - Q_k = Q_k^\perp} \; . \tag{1.201}$$

Note that Q_k and $Q_k^\perp$ are **orthogonal projections**.
Transposing (1.201), we obtain

$$(E_{k+1}^*)^T \;=\; Q_k \; (E_k^*)^T \;+\; Q_k^\perp \; D_{k+1}^T \; . \tag{1.202}$$

If we apply the affine covariant Lipschitz condition to $D_{k+1}\ \delta x^k$, from (1.201) we get the estimate

$$\frac{\|E_{k+1}^* \delta x^k\|}{\|\delta x^k\|} \ = \ \frac{\|D_{k+1} \delta x^k\|}{\|\delta x^k\|} \ \leq \ \bar{t}_k \ , \tag{1.203}$$

whereas (1.202) results in

$$\|E_{k+1}^*\| \ = \ \max_{v \neq 0} \frac{\|(E_{k+1}^*)^T v\|}{\|v\|} \ \leq \tag{1.204}$$

$$\leq \ \max_{v \neq 0} \frac{\|(E_k^*)^T v\|}{\|v\|} \ + \ \max_{v \neq 0} \frac{|(D_{k+1} \delta x^k, v)|}{\|\delta x^k\| \ \|v\|} \ \leq \ \|E_k^*\| \ + \ \bar{t}_k \ .$$

In view of (1.199), assuming **uniform boundedness**

$$\overline{\Theta}_k \ < \ 1 \quad , \quad \|E_k^*\| \leq \overline{\eta} \quad , \quad k \in \mathbb{N}_0 \ ,$$

it is natural to assume

$$\overline{\Theta}_k \ \leq \ \frac{\overline{\eta} \ + \ \bar{t}_0}{1 - \overline{\eta}} \ := \ \overline{\Theta} \ . \tag{1.205}$$

This gives

$$\|E_{k+1}^*\| \ \leq \ \|E_0^*\| \ + \ \sum_{\ell=0}^{k} \bar{t}_\ell \ < \ \|E_0^*\| \ + \ \frac{\bar{t}_0}{1 - \overline{\Theta}} \ , \tag{1.206}$$

so that we may set

$$\overline{\eta} \ := \ \|E_0^*\| \ + \ \frac{\bar{t}_0}{1 - \overline{\Theta}} \ . \tag{1.207}$$

If we insert the expression for $\overline{\eta}$ into (1.205) and solve for $\bar{t}_0$, we obtain

$$2 \ \bar{t}_0 \ = \ (1 - \overline{\Theta}) \ \Big(\overline{\Theta} \ - \ (1 + \overline{\Theta}) \ \|E_0^*\| \Big) \ .$$

Taking into account that

$$\bar{t}_0 \ = \ \frac{1}{2} \ (t_0 \ + \ t_1) \ \leq \ \frac{1}{2} \ (1 \ + \ \overline{\Theta}) \ t_0 \ ,$$

leads to the requirement

$$t_0 \ \leq \ (1 - \overline{\Theta}) \ \Big(\frac{\overline{\Theta}}{1 + \overline{\Theta}} \ - \ \|E_0^*\| \Big) \ . \tag{1.208}$$

Hence, $\|E_0^*\|$ has to satisfy

$$\|E_0^*\| \;<\; \frac{\overline{\Theta}}{1+\overline{\Theta}} \;<\; \frac{1}{2} \quad \text{for } \overline{\Theta} < 1 \;. \tag{1.209}$$

On the other hand,

$$\|E_0^*\| \;=\; \|F'(x^*)^{-1}J_0 - I\| \;=\; \|F'(x^*)^{-1}\Big(J_0 - F'(x^*) \pm F'(x^0)\Big)\| \;\leq$$

$$\leq \underbrace{\|F'(x^*)^{-1}\Big(J_0 - F'(x^0)\Big)\|}_{=\,\delta_0} + \underbrace{\|F'(x^*)^{-1}\Big(F'(x^0) - F'(x^*)\Big)\|}_{\leq\, \omega\|x^0-x^*\|=t_0} \;.$$

Replacing $\|E_0^*\|$ in (1.208) by $t_0 + \delta_0$ gives

$$(2-\overline{\Theta})\; t_0 \;\leq\; (1-\overline{\Theta})\;\Big(\frac{\overline{\Theta}}{1+\overline{\Theta}} \;-\; \delta_0\Big) \quad \Longrightarrow$$

$$t_0 \;\leq\; \frac{1-\overline{\Theta}}{2-\overline{\Theta}}\;\Big(\frac{\overline{\Theta}}{1+\overline{\Theta}} \;-\; \delta_0\Big)\;,$$

so that (1.208) can be replaced by the assumptions (1.188) and (1.189).
Now, for a $\overline{\Theta} < 1$ satisfying (1.188) and (1.189), the **linear convergence** (1.190) follows from (1.199), whereas (1.191) follows by inserting $\overline{\eta}$ into (1.197).
Finally, the **bounded deterioration property** (1.193) results from the insertion of (1.189) into (1.206).
What remains to be proved is the **superlinear convergence** (1.192). In view of (1.201), we have

$$\|(E_{k+1}^*)^T v\|^2 \;=\; \|(E_k^*)^T v\|^2 \;-\; \frac{(E_k^*\delta x^k, v)^2}{\|\delta x^k\|^2} \;+\; \frac{(D_{k+1}\delta x^k, v)^2}{\|\delta x^k\|^2} \;.$$

Summing over all $0 \leq k \leq \ell$, it follows that

$$\sum_{k=0}^{\ell} \frac{(E_k^*\delta x^k, v)^2}{\|v\|^2\;\|\delta x^k\|^2} \;=\; \frac{\|E_0^* v\|^2}{\|v\|^2} \;-\; \frac{\|E_{\ell+1}^* v\|^2}{\|v\|^2} \;+\; \sum_{k=0}^{\ell} \frac{(D_{k+1}\delta x^k, v)^2}{\|v\|^2\;\|\delta x^k\|^2} \;.$$

Now, letting $\ell \to \infty$ and using (1.203), i.e.,

$$\frac{\|E_{k+1}^*\delta x^k\|}{\|\delta x^k\|} \;=\; \frac{\|D_{k+1}\delta x^k\|}{\|\delta x^k\|} \;\leq\; \bar{t}_k \;,$$

we obtain

$$\sum_{k=0}^{\infty} \frac{(E_k^* \delta x^k, v)^2}{\|v\|^2 \, \|\delta x^k\|^2} \;\leq\; \|E_0^*\|^2 \;+\; \sum_{k=0}^{\infty} \bar{t}_k^2 \;\leq\; \|E_0^*\|^2 \;+\; \sum_{k=0}^{\infty} \overline{\Theta}^{2k} \, \bar{t}_0^2 \; .$$

Since $\bar{t}_0^2 \;\leq\; \frac{1}{2}(1+\overline{\Theta}^2) t_0^2$, it follows that

$$\sum_{k=0}^{\infty} \frac{(E_k^* \delta x^k, v)^2}{\|v\|^2 \, \|\delta x^k\|^2} \;\leq\; \|E_0^*\|^2 \;+\; \frac{1}{2} \, \frac{1+\overline{\Theta}^2}{1-\overline{\Theta}^2} \, t_0^2 \; .$$

the right-hand side in the previous inequality is bounded, whence

$$\lim_{k \to \infty} \frac{(E_k^* \delta x^k, v)^2}{\|v\|^2 \, \|\delta x^k\|^2} \;=\; 0 \quad , \quad v \in \mathbb{R}^n \setminus \{0\} \; .$$

Consequently, setting

$$\xi_k \;:=\; \frac{\delta x^k}{\|\delta x^k\|} \; ,$$

we have

$$\lim_{k \to \infty} E_k^* \xi_k \;=\; 0 \; ,$$

which proves (1.194).
Finally, using the previous result in (1.197) proves superlinear convergence.

•

4.2.2 Algorithmic aspects of the affine covariant Quasi-Newton method

(i) Recursive Good Broyden algorithm

The recursion (1.184) cannot be used directly for the computation of the Quasi-Newton increments. To come up with a computationally feasible recursion, we rewrite (1.184) according to

$$J_{k+1}^{-1} \;=\; \left(I \;+\; \frac{\delta x^{k+1} (\delta x^k)^T}{\|\delta x^k\|^2} \right) J_k^{-1} \; . \tag{1.210}$$

This leads to the following **product representation**

$$J_{k+1}^{-1} \;=\; \prod_{\ell=0}^{k-1} \left(I \;+\; \frac{\delta x^{\ell+1} (\delta x^\ell)^T}{\|\delta x^\ell\|^2} \right) J_0^{-1} \; , \tag{1.211}$$

which can be efficiently used in actual computations.

(ii) Condition number monitor

In order to monitor the condition number of the approximations of the Jacobians, we provide the following elementary result for rank 1 matrices.

Lemma 4.1 Condition number of rank 1 matrices

For a rank 1 matrix A of the form

$$A \;=\; I \;-\; \frac{uv^T}{v^T v} \quad , \quad \Theta \;:=\; \frac{\|u\|}{\|v\|} \;<\; 1 \;,$$

the condition number can be bounded according to

$$\mathrm{cond}(A) \;\leq\; \frac{1+\Theta}{1-\Theta} \;.$$

Proof. The assertion readily follows from the two estimates

$$\|A\| \;\leq\; 1 \;+\; \|\frac{uv^T}{v^T v}\| \;\leq\; 1+\Theta$$

and

$$\|A^{-1}\| \;\leq\; \Big(1 \;-\; \|\frac{uv^T}{v^T v}\|\Big)^{-1} \;\leq\; \frac{1}{1-\Theta} \;.$$

•

Applying Lemma 4.1 to the recursions (1.56) and (1.60) and observing

$$\Theta_k \;=\; \frac{\|\bar{\delta}x^{k+1}\|}{\|\delta x^k\|} \;<\; \frac{1}{2} \;,$$

we get the **condition number estimate**

$$\mathrm{cond}(J_{k+1}) \;\leq\; \frac{1+\Theta_k}{1-\Theta_k}\,\mathrm{cond}(J_k) \;<\; 3\,\mathrm{cond}(J_k) \;. \tag{1.212}$$

(iii) Convergence monitor

In accordance with the result of the condition number estimate in **(ii)** we monitor the convergence by

$$\Theta_k \;<\; \frac{1}{2} \;, \tag{1.213}$$

i.e., if $\Theta_k \geq \frac{1}{2}$, no convergence is detected.

(iv) Termination criterion

Denoting by $XTOL$ a user specified accuracy, the Quasi-Newton iteration will be terminated, if

$$\|\delta x^{k+1}\| \;\leq\; XTOL\;. \tag{1.214}$$

4.3 Affine contravariant Quasi-Newton method

Affine contravariant Quasi-Newton methods require to reformulate the secant condition (1.164) in an affine contravariant manner. For that purpose, setting $F^{k+1} := F(x^{k+1}) - F(x^k)$, we rewrite (1.164) according to

$$\underbrace{x^{k+1} - x^k}_{=\,\delta x^k} \;=\; J^{-1}\delta F^{k+1}\;.$$

multiplication by J_k results in

$$\underbrace{J_k\delta x^k}_{=\,-F(x^k)} \;=\; J_k J^{-1}\delta F^{k+1}\;,$$

whence

$$J_k J^{-1}\delta F^{k+1} \;=\; \delta F^{k+1} \;-\; F(x^{k+1})\;.$$

We thus get

$$\underbrace{\left(I - J_k J^{-1}\right)}_{=\,E_k(J)}\,\delta F^{k+1} \;=\; F(x^{k+1})\;. \tag{1.215}$$

We note that any Jacobian rank 1 update of the form

$$J_{k+1}^{-1} \;=\; J_k^{-1}\left(I \;-\; \frac{F(x^{k+1})v^T}{v^T\delta F^{k+1}}\right) \quad,\quad v\in\mathbb{R}^n\setminus\{0\}$$

satisfies the **affine contravariant secant condition** (1.215).
In particular, for $v = \delta F^{k+1}$ we recover the **Bad Broyden update** (1.176).

4.3.1 Affine contravariant Quasi-Newton convergence theory

We begin with some useful properties of the Bad Broyden update.

Theorem 4.3 Properties of the affine contravariant Quasi-Newton method

For Broyden's affine contravariant rank 1 update (Bad Broyden)

$$J_{k+1}^{-1} \;=\; J_k^{-1}\left(I \;-\; \frac{\overline{\delta}F(x^{k+1})(\delta F^{k+1})^T}{\|\delta F^{k+1}\|^2}\right) \tag{1.216}$$

assume **local residual contraction**

$$\Theta_k \;=\; \frac{\|F(x^{k+1})\|}{\|F(x^k)\|} \;<\; 1\;. \tag{1.217}$$

Then, there holds:

(i) The update matrix J_{k+1} is a **least change update** in the sense that

$$\|E_k(J_{k+1})\| \;\leq\; \|E_k(J)\| \quad,\quad J \in \mathcal{S}_{k+1}\;, \tag{1.218}$$

$$\|E_k(J_{k+1})\| \;\leq\; \frac{\Theta_k}{1-\Theta_k}\;. \tag{1.219}$$

(ii) If J_k is regular, then J_{k+1} is regular as well. J_{k+1} can be represented according to

$$J_{k+1} \;=\; \left(I \;-\; \frac{F(x^{k+1})(\delta F^{k+1})^T}{(\delta F^{k+1})^T F(x^k)}\right) J_k\;. \tag{1.220}$$

(iii) The **Quasi-Newton increment** δx^{k+1} is given by

$$\delta x^{k+1} \;=\; -\,J_{k+1}^{-1}F(x^{k+1}) \;=\; \left(1 \;-\; \frac{(\delta F^{k+1})^T F(x^{k+1})}{\|\delta F^{k+1}\|^2}\right) \underbrace{\left(-\,J_k^{-1}F(x^{k+1})\right)}_{=:\;\overline{\delta}x^{k+1}} \tag{1.221}$$

Proof. For the proof of (1.218),(1.219) we have

$$E_k(J_{k+1}) \;=\; I \;-\; J_k J_{k+1}^{-1} \;=\; \frac{F(x^{k+1})(\delta F^{k+1})^T}{\|\delta F^{k+1}\|^2}\;,$$

which gives

$$\|E_k(J_{k+1})\| \;=\; \frac{\|E_k(J_{k+1})\delta F^{k+1}\|}{\|\delta F^{k+1}\|} \;=\; \frac{\|F(x^{k+1})\|}{\|\delta F^{k+1}\|} \;=$$

$$=\; \frac{\|E_k(J)\delta F^{k+1}\|}{\|\delta F^{k+1}\|} \;\leq\; \|E_k(J)\| \quad,\quad J \in \mathcal{S}_{k+1}\;.$$

Using

$$\|F(x^{k+1}) - F(x^k)\| \;\geq\; \|F(x^k)\| \;-\; \|F(x^{k+1})\| \;=\; \|f(x^k)\|\left(1 \;-\; \frac{\|F(x^{k+1})\|}{\|F(x^k)\|}\right),$$

we find

$$\|E_k(J_{k+1})\| \;=\; \frac{\|F(x^{k+1})\|}{\|\delta F^{k+1}\|} \;\leq\; \frac{\Theta_k}{1-\Theta_k}\;.$$

•

In the the subsequent convergence proof for the affine contravariant Quasi-Newton method we need the following technical result:

Lemma 4.2 An elementary technical estimate

Assume $0 < \Theta < 1$, $0 \leq \eta_0 < \Theta$ and

$$t \;\leq\; \frac{\Theta \;-\; \eta_0}{1 + \eta_0 + \frac{4}{3}(1-\Theta)^{-1}} \;.$$

Setting

$$\eta \;=\; \eta_0 \;+\; \frac{t}{(1-t)(1-\Theta)} \;,$$

there holds

$$\eta \;+\; (1+\eta)\, t \;\leq\; \Theta \;.$$

Proof. For t we have

$$t \;\leq\; \frac{\Theta}{1 + \frac{4}{3(1-\Theta}\,)} \;=\; \frac{3\Theta(1-\Theta)}{7 - 3\Theta} \;=:\; g(\Theta) \;.$$

The function $g(\Theta)$ attains its maximum in $\Theta^* = \frac{7-\sqrt{28}}{3}$ with $g(\Theta^*) < \frac{1}{7}$. Hence,

$$\Theta \;\geq\; \eta_0 \;+\; (1 + \eta_0 + \frac{\frac{4}{3}}{1-\Theta}\,)\; t \;=$$

$$=\; \eta_0 \;+\; \frac{\frac{7}{6}t}{1-\Theta} \;+\; (1 + \eta_0 + \frac{\frac{1}{6}}{1-\Theta}\,)\; t \;\geq$$

$$\geq\; \eta_0 \;+\; \frac{t}{(1-t)(1-\Theta)} \;+\; (1 + \eta_0 + \frac{t}{(1-t)(1-\Theta)}\,)\; t \;=\; \eta \;+\; (1+\eta)t \;.$$

•

Theorem 4.4 Convergence of the affine contravariant Quasi-Newton method

Suppose that that $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$, $D \subset \mathbb{R}^n$ convex, is continuously differentiable on D and let $x^* \in D$ be the unique solution of $F(x) = 0$ in D with invertible Jacobian $F'(x^*)$. Assume that the following **affine contravariant Lipschitz condition** is satisfied

$$\| \Big(F'(x) - F'(x^+) \Big) \, (y - x) \| \;\leq\; \omega \, \| F'(x^*)(x - x^*) \| \, \| F'(x^*)(y - x) \| \tag{1.222}$$

where $x, y \in D$, and denote by

$$E_k^* \;:=\; I \;-\; F'(x^*) J_k^{-1} \tag{1.223}$$

the **affine contravariant deterioration matrix**.
For some $0 < \overline{\Theta} < 1$ assume further that:

(a) The **initial approximate Jacobian** J_0 satisfies

$$\overline{\eta}_0 \;:=\; \| E_0^* \| \;<\; \overline{\Theta} \, . \tag{1.224}$$

(b) The **initial guess** $x^0 \in D$ satisfies

$$t_0 \;:=\; \omega \, \| F'(x^*)(x^0 - x^*) \| \;\leq\; \frac{\overline{\Theta} - \overline{\eta}_0}{1 + \overline{\eta}_0 + \frac{4}{3}(1 - \overline{\Theta})^{-1}} \, . \tag{1.225}$$

Then, there holds:

(i) The **Quasi-Newton iterates** $x^k, k \in \mathbb{N}_0$ converge to x^* according to

$$\| F'(x^*)(x^{k+1} \;-\; x^*) \| \;\leq\; \overline{\Theta} \, \| F'(x^*)(x^k \;-\; x^*) \| \, , \tag{1.226}$$

$$\| F(x^{k+1}) \| \;\leq\; \overline{\Theta} \, \| F(x^k) \| \, . \tag{1.227}$$

We have **superlinear convergence** in the sense that

$$\lim_{k \to \infty} \frac{\| F(x^{k+1}) \|}{\| F(x^k) \|} \;=\; 0 \, . \tag{1.228}$$

(ii) The following **affine contravariant bounded deterioration property** holds true

$$\| E_k^* \| \;\leq\; \overline{\eta}_0 \;+\; \frac{t_0}{(1 - t_0)(1 - \overline{\Theta})} \;\leq\; \overline{\Theta} \, . \tag{1.229}$$

Moreover, asymptotically we have

$$\lim_{k\to\infty} \frac{\|E_k^* \delta F^{k+1}\|}{\|\delta F^{k+1}\|} \;=\; 0 \;. \tag{1.230}$$

Proof. We first derive an **estimate for the iterative residuals** $F(x^{k+1})$. We have

$$F(x^{k+1}) \;=\; \underbrace{F(x^k)}_{=\,-J_k\delta x^k} \;+\; \int_0^1 F'(x^k + t\delta x^k)\;\delta x^k\;dt \;=$$

$$=\; \int_0^1 \Big(F'(x^k + t\delta x^k) \;-\; F'(x^*)\Big)\;\delta x^k\;dt \;+\; \underbrace{\Big(F'(x^*) -\; J_k\Big)}_{=\,(F'(x^*)J_k^{-1} - I)\;\underbrace{J_k\delta x^k}_{=\,-F(x^k)}} \quad \delta x^k \;.$$

Using the **affine contravariant Lipschitz condition** (1.222), it follows that

$$\|F(x^{k+1})\| \;\leq$$

$$\leq\; \int_0^1 \|\Big(F'(x^k + t\delta x^k) \;-\; F'(x^*)\Big)\;\delta x^k\|\;dt \;+\; \|\underbrace{\Big(F'(x^*)J_k^{-1} \;-\; I\Big)}_{=\,-E_k^*}\;F(x^k)\| \;\leq$$

$$\leq\; \omega \int_0^1 \|F'(x^*) \underbrace{(x^k + t\delta x^k - x^*)}_{=\,(1-t)(x^k-x^*)+t(x^{k+1}-x^*)} \| \;\|F'(x^*)\delta x^k\|\;dt \;+\; \|E_k^* F(x^k)\| \;\leq$$

$$\leq\; \omega \int_0^1 \Big((1-t)\|F'(x^*)(x^k - x^*)\| \;+\; t\|F'(x^*)(x^{k+1} - x^*)\|\Big)\;\|F'(x^*)\delta x^k\|\;dt \;+$$

$$+\; \|E_k^*\|\;\|F(x^k)\| \;=\; \frac{1}{2}\;(t_k + t_{k+1})\;\|\; \underbrace{F'(x^*)\delta x^k}_{=\,(I-E_k^*)\;\underbrace{J_k\delta x^k}_{=\,-F(x^k)}} \;\| \;+\; \|E_k^*\|\;\|F(x^k)\|\;,$$

where

$$t_k \;:=\; \omega\, \|F'(x^*)(x^k - x^*)\| \quad , \quad k \in \mathbb{N}_0 \; .$$

Setting

$$\bar{t}_k \;:=\; \frac{1}{2}\,(t_k + t_{k+1}) \quad , \quad k \in \mathbb{N}_0 \; ,$$

we obtain

$$\|F(x^{k+1})\| \;\leq\; \bar{t}_k \|(E_k^* - I)F(x^k)\| \;+\; \|E_k^*\|\, \|F(x^k)\| \leq \tag{1.231}$$

$$\leq\; \left[\bar{t}_k(1 + \|E_k^*\|) \;+\; \|E_k^*\|\right] \|F(x^k)\| \; .$$

Likewise, for the **iterative error** $F'(x^*)(x^{k+1} - x^*)$ we get

$$F'(x^*)(x^{k+1} - x^*) \;=\; F'(x^*)(x^k - x^*) \;+\; \underbrace{F'(x^*) \underbrace{\delta x^k}_{= -J_k^{-1}F(x^k)}}_{= (E_k^* - I)F(x^k)} \;=$$

$$=\; F'(x^*)(x^k - x^*) \;+\; \underbrace{F(x^*) \;-\; F(x^k)}_{= \int_0^1 F'(x^* + t(x^* - x^k))(x^* - x^k)dt} \;+\; E_k^* F(x^k) \; ,$$

and hence,

$$\|F'(x^*)(x^k - x^*)\| \;\leq$$

$$\leq\; \omega \int_0^1 t\, \|F'(x^*)(x^k - x^*)\|^2 \; dt \;+\; \|E_k^*\|\, \|F(x^k) \;\pm F'(x^*)(x^k - x^*)\| \;\leq$$

$$\leq\; \frac{\omega}{2}\, \|F'(x^*)(x^k - x^*)\|^2 \;+\; \|E_k^*\| \left(\|F'(x^*)(x^k - x^*) - F(x^k)\| \;+\right.$$

$$\left. +\; \|F'(x^*)(x^k - x^*)\|\right) \; .$$

Treating $F'(x^*)(x^k - x^*) - F(x^k)$ as before and multiplying by ω yields

$$t_{k+1} \;\leq\; \frac{1}{2}\, t_k^2 \;+\; \|E_k^*\|\, (\frac{1}{2}\, t_k^2 \;+\; t_k) \;= \tag{1.232}$$

$$= \left(\|E_k^*\| \; + \; \frac{1 + \|E_k^*\|}{2} \, t_k \right) t_k \; .$$

We further study the **approximation properties of the Jacobian updates**. For the **deterioration matrix** E_{k+1}^* we obtain

$$E_{k+1}^* \; = \; I \; - \; F'(x^*) J_{k+1}^{-1} \; = \; I \; - \; F'(x^*) J_k^{-1} \left(I \; - \; \frac{F(x^{k+1})(\delta F^{k+1})^T}{\|\delta F^{k+1}\|^2} \right) \; =$$

$$= (I - F'(x^*) J_k^{-1}) \left(I - \frac{(F(x^{k+1}) \pm F(x^k))(\delta F^{k+1})^T}{\|\delta F^{k+1}\|^2} \right) + \frac{F(x^{k+1})(\delta F^{k+1})^T}{\|\delta F^{k+1}\|^2} =$$

$$= \underbrace{(I - F'(x^*) J_k^{-1})}_{= E_k^*} \underbrace{(I \; - \; \frac{\delta F^{k+1}(\delta F^{k+1})^T}{\|\delta F^{k+1}\|^2})}_{= (I - Q_k) \; = \; Q_k^\perp} \; -$$

$$- \; (I - F'(x^*) J_k^{-1}) \; \frac{(F(x^k) \pm F(x^{k+1}))(\delta F^{k+1})^T}{\|\delta F^{k+1}\|^2} \; + \; \frac{F(x^{k+1})(\delta F^{k+1})^T}{\|\delta F^{k+1}\|^2} \; =$$

$$= \; E_k^* Q_k^\perp \; + \; F'(x^*) J_k^{-1} \Big(\frac{F(x^{k+1})(\delta F^{k+1})^T}{\|\delta F^{k+1}\|^2} \; - \; \underbrace{\frac{\delta F^{k+1}(\delta F^{k+1})^T}{\|\delta F^{k+1}\|^2}}_{= \; Q_k} \Big) \; .$$

Finally, taking advantage of

$$\frac{F(x^{k+1})(\delta F^{k+1})^T}{\|\delta F^{k+1}\|^2} \; Q_k \; = \; \frac{F(x^{k+1})(\delta F^{k+1})^T}{\|\delta F^{k+1}\|^2} \; ,$$

we obtain

$$E_{k+1}^* \; = \; E_k^* Q_k^\perp \; + \; F'(x^*) J_k^{-1} \Big(\frac{F(x^{k+1})(\delta F^{k+1})^T}{\|\delta F^{k+1}\|^2} \; Q_k \; - \; Q_k \Big) \; = \quad (1.233)$$

$$= \; E_k^* Q_k^\perp \; + \; \Big(I \; - \; F'(x^*) \underbrace{J_k^{-1} \left(I \; - \; \frac{F(x^{k+1})(\delta F^{k+1})^T}{\|\delta F^{k+1}\|^2} \right)}_{= \; J_{k+1}^{-1}} \Big) \; Q_k \; =$$

$$= \; E_k^* Q_k^\perp \; + \; E_{k+1}^* \; Q_k \; .$$

The representation (1.233) of E^*_{k+1} readily yields the estimate

$$\|E^*_{k+1}\| \;\leq\; \|E^*_k Q^\perp_k\| \;+\|E^*_{k+1}Q_k\| \;\leq \tag{1.234}$$

$$\leq\; \|E^*_k\| \;+\; \frac{\|E^*_{k+1}\delta F^{k+1}\|}{\|\delta F^{k+1}\|}\;.$$

Recalling the definition (1.223) of E^*_{k+1},, we have

$$E^*_{k+1}\delta F^{k+1} \;=\; \delta F^{k+1} \;-\; F'(x^*)\; J^{-1}_{k+1}\; \delta F^{k+1}\;.$$

On the other hand, the update formula (**??**) gives

$$J^{-1}_{k+1}\; \delta F^{k+1} \;=\; J^{-1}_k \delta F^{k+1} \;-\; J^{-1}_k F(x^{k+1}) \;=\; -\; J^{-1}_k F(x^k) \;=\; \delta x^k\;,$$

and hence

$$E^*_{k+1}\delta F^{k+1} \;=\; \delta F^{k+1} \;-\; F'(x^*)\;\delta x^k \;= \tag{1.235}$$

$$= \int_0^1 \Big(F'(x^k + t\delta x^k) \;-\; F'(x^*)\Big)\;\delta x^k\; dt\;.$$

By the **affine contravariant Lipschitz condition** and using (1.235)

$$\|E^*_{k+1}\delta F^{k+1}\| \;\leq\; \omega \int_0^1 \|F'(x^*)(x^k + t\delta x^k - x^*)\|\;\|F'(x^*)\delta x^k\|\; dt \;\leq$$

$$\leq \omega \int_0^1 \Big[t\|F'(x^*)(x^{k+1} - x^*)\| + (1-t)\|F'(x^*)(x^k - x^*)\|\Big]\|F'(x^*)\delta x^k\|\; dt \leq$$

$$\leq\; \bar{t}_k\;\|F'(x^*)\delta x^k\| \;=\; \bar{t}_k\;\|E_{k+1}\delta F^{k+1} - \delta F^{k+1}\| \;\leq\; \|E_{k+1}\delta F^{k+1}\| + \|\delta F^{k+1}\|\;,$$

whence

$$\|E^*_{k+1}\delta F^{k+1}\| \;\leq\; \frac{\bar{t}_k}{1-\bar{t}_k}\;\|\delta F^{k+1}\|\;.$$

Using the previous estimate in (1.234) results in

$$\|E^*_{k+1}\| \;\leq\; \|E^*_k\| \;+\; \frac{\bar{t}_k}{1-\bar{t}_k}\;. \tag{1.236}$$

The **bounded deterioration property** and the **contraction of the residuals** is now proved by an induction argument. We assume that we have

$$\|E_k^*\| \;\leq\; \|E_0^*\| \;+\; \frac{\sum_{\ell=0}^{k} \overline{\Theta}^{\ell}\; t_0}{1-t_0} \;\leq\; \overline{\eta} \;:=\; \|E_0^*\| \;+\; \frac{t_0}{(1-t_0)(1-\overline{\Theta})} \;.$$

and

$$t_k \;\leq\; \overline{\Theta}^k \; t_0 \;.$$

Then, (1.232) in combination with Lemma 4.2 yields

$$t_{k+1} \;\leq\; (\overline{\eta} + (1+\overline{\eta})t_0)t_k \;\leq\; \overline{\Theta}\; t_k \;\leq\; \overline{\Theta}^k t_0 \;.$$

Moreover, (1.236) gives us

$$\|E_{k+1}^*\| \;\leq\; \|E_k^*\| \;+\; \frac{t_k}{1-t_0} \;\leq$$

$$\leq\; \|E_0^*\| \;+\; \frac{\sum_{\ell=0}^{k-1} \overline{\Theta}^{\ell} t_0}{1-t_0} \;+\; \frac{\overline{\Theta}^{k+1} t_0}{1-t_0} \;\leq$$

$$\leq\; \|E_0^*\| \;+\; \frac{\sum_{\ell=0}^{k} \overline{\Theta}^{\ell} t_0}{1-t_0} \leq\; \overline{\eta} \;.$$

In summary, induction on k shows the **bounded deterioration property**

$$\|E_k^*\| \;\leq\; \overline{\Theta}$$

as well as

$$t_{k+1} \;\leq\; t_k \;,$$

which in view of (1.232) and Lemma 4.2 implies **contraction of the residuals**

$$\|F(x^{k+1})\| \;\leq\; \overline{\Theta}\; \|F(x^k)\| \;.$$

We finally prove **superlinear convergence**. For that purpose, we use the orthogonal decomposition (1.233) in the form

$$(E_{k+1}^*)^T \;=\; Q_k^{\perp}(E_k^*)^T \;+\; Q_k(E_{k+1}^*)^T \;.$$

Observing

$$Q_k(E_{k+1}^*)^T v \;=\; \delta F^{k+1}\,\frac{(\delta F^{k+1},(E_{k+1}^*)^T v)}{\|\delta F^{k+1}\|^2} \;=\; \delta F^{k+1}\,\frac{(E_{k+1}^*\delta F^{k+1},v)}{\|\delta F^{k+1}\|^2}\;,$$

we get

$$\begin{aligned}\|(E_{k+1}^*)^T v\|^2 \;&=\; \|Q_k^\perp (E_k^*)^T v\|^2 \;+\; \|Q_k(E_{k+1}^*)^T v\|^2 \;=\\ &=\; \|(E_k^*)^T v\|^2 \;-\; \|Q_k^\perp (E_k^*)^T v\|^2 \;+\; \|Q_k(E_{k+1}^*)^T v\|^2 \;=\\ &=\; \|(E_k^*)^T v\|^2 \;-\; \frac{(E_k^*\delta F^{k+1},v)^2}{\|\delta F^{k+1}\|^2} \;+\; \frac{(E_{k+1}^*\delta F^{k+1},v)^2}{\|\delta F^{k+1}\|^2}\;.\end{aligned}$$

Summing over all $0 \le k \le \ell$ yields

$$\sum_{k=0}^{\ell}\frac{(E_k^*\delta F^{k+1},v)^2}{\|\delta F^{k+1}\|^2\|v\|^2} \;=\; \frac{\|E_0^*\|^2}{\|v\|^2} \;-\; \frac{\|E_{\ell+1}^*\|^2}{\|v\|^2} \;+\; \sum_{k=0}^{\ell}\frac{(E_{k+1}^*\delta F^{k+1},v)^2}{\|\delta F^{k+1}\|^2\|v\|^2}\;.$$

Estimating from above by neglecting the negative term and observing (1.236), for $\ell \to \infty$, we obtain

$$\sum_{k=0}^{\ell}\frac{(E_k^*\delta F^{k+1},v)^2}{\|\delta F^{k+1}\|^2\|v\|^2} \le \|E_0^*\|^2 + \sum_{k=0}^{\ell}\left(\frac{t_k}{1-t_k}\right)^2 \le \|E_0^*\|^2 + \frac{t_0^2}{(1-t_0)^2}\underbrace{\sum_{k=0}^{\infty}\overline{\Theta}^{2k}}_{=\,\frac{1}{1-\overline{\Theta}^2}}\;.$$

Due to the boundedness of the right-hand side, we must have

$$\lim_{k\to\infty}\frac{(E_k^*\delta F^{k+1},v)^2}{\|\delta F^{k+1}\|^2\|v\|^2} \;=\; 0\;,$$

and thus

$$\lim_{k\to\infty}\|E_k^*\| \;=\; 0\;.$$

The **superlinear convergence** of the residuals results from the following reasoning: Observing

$$E_k^* F(x^{k+1}) \;=\; F(x^{k+1}) \;-\; F'(x^*)J_k^{-1}F(x^{k+1})\;,$$

we have

$$\|F(x^{k+1})\| \;-\; \|F'(x^*)J_k^{-1}F(x^{k+1})\| \;\le\; \|E_k^*F(x^{k+1})\| \;\le\; \|E_k^*\|\|F(x^{k+1})\|\;,$$

and hence,

$$\|F(x^{k+1})\| \;\leq\; \frac{\|F'(x^*)J_k^{-1}F(x^{k+1})\|}{1-\|E_k^*\|}\;. \tag{1.237}$$

On the other hand, in view of the update formula (1.216)

$$F'(x^*)J_k^{-1}F(x^{k+1}) \;=\; (E_{k+1}^*-E_k^*)\delta F^{k+1}\;,$$

which gives

$$\|F'(x^*)J_k^{-1}F(x^{k+1})\| \;\leq\; \underbrace{\|E_{k+1}^*\|}_{\leq\, \|E_k^*\|+\frac{\bar{t}_k}{1-\bar{t}_k}} \|\delta F^{k+1}\| \;+\; \|E_k^*\|\|\delta F^{k+1}\|\;.$$

Taking into account that

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

we arrive at

$$\|F'(x^*)J_k^{-1}F(x^{k+1})\| \;\leq\; \left(2\|E_k^*\| \;+\; \frac{\bar{t}_k}{1-\bar{t}_k}\right)(1+\bar{\theta})\;\|F(x^k)\|\;.$$

Using the previous estimate in (1.237) yields

$$\|F(x^{k+1})\| \;\leq\; \frac{\left(2\|E_k^*\| \;+\; \frac{\bar{t}_k}{1-\bar{t}_k}\right)}{1-\|E_k^*\|}\;\|F(x^k)\|\;. \tag{1.238}$$

Since $\|E_k^*\| \to 0$ and $\bar{t}_k \to 0$ as $k \to \infty$, (1.238) proves superlinear convergence of the residuals.

•

4.3.2 Algorithmic aspects of the affine contravariant Quasi-Newton method

(i) Condition number monitor

As in the affine covariant Quasi-Newton method, we track the condition number of the Jacobian rank 1 updates. For $\Theta_k < \frac{1}{2}$, an application of Lemma 4.1 yields

$$\mathrm{cond}(J_{k+1}) \;\leq\; \underbrace{\mathrm{cond}\left(I-\frac{F(x^{k+1})(\delta F^{k+1})^T}{\|\delta F^{k+1}\|^2}\right)}_{\leq\, \frac{1}{1-2\Theta_k}}\;\mathrm{cond}(J_k)\;.$$

(ii) Convergence monitor

We choose

$$\Theta_{max} \;<\; \frac{1}{2} \quad , \quad \text{e.g.} \quad \Theta_{max} \;=\; \frac{1}{4}$$

and check

$$\Theta_k \;\leq\; \Theta_{max} \;. \tag{1.239}$$

If (1.239) is violated, we stop the algorithm (no convergence).

(iii) Termination criterion

Given a user specified tolerance FTOL, the Quasi-Newton iteration will be stopped, if

$$\|F(x^k)\| \;\leq\; \text{FTOL} \;. \tag{1.240}$$

5. Global Newton methods

The convergence results of the previous chapters stated **local convergence** of Newton or Newton-like methods under a restriction on the initial guess in terms of the initial Kantorovich quantity. If this condition is not satisfied, there is no guaranteed convergence and **globalization strategies** have to be imposed such as **steepest descent** or **trust region methods**.
In the sequel, we derive a globalization strategy that is based on an appropriate **damping of the Newton increments** within the three affine invariance classes and design related **residual** and **error-oriented monotonicity tests** as well as a **convex functional test** along with **adaptive trust region strategies** for a suitable selection of the damping parameter.

5.1 The Newton path

A widely used globalization concept relies on a decrease of the **residual level function**

$$T(x) \;:=\; \frac{1}{2}\,\|F(x)\|^2 \;=\; \frac{1}{2}\,F(x)^T F(x) \tag{1.241}$$

in the sense that we require the **monotonicity criterion**

$$T(x^{k+1}) \;<\; T(x^k) \quad , \quad \text{if} \quad T(x^k) \neq 0\;. \tag{1.242}$$

We associate with the residual level function T the **level set**

$$G(z) \;:=\; \{\; x \in D \subset \mathbb{R}^n \mid T(x) \leq T(z) \;\}\;. \tag{1.243}$$

In terms of the level set G, the monotonicity criterion (1.242) can be stated as

$$x^{k+1} \in \text{int}\; G(x^k) \quad , \quad \text{if} \quad \text{int}\; G(x^k) \neq \emptyset\;. \tag{1.244}$$

In the **steepest descent method**, the gradient of the level function is used as the direction of the iterative correction

$$\begin{aligned} \Delta x^k \;&=\; -\;\text{grad}\; T(x^k) \;=\; -\;F'(x^k)F(x^k)\;, \\ x^{k+1} \;&=\; x^k \;+\; s_k\;\Delta x^k\;, \end{aligned} \tag{1.245}$$

where $s_k > 0$ is an appropriate **steplength parameter**.
The convergence of the steepest descent method is assured by the following result.

Lemma 5.1 Downhill property

Assume that $F : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$ is continuously differentiable on D and $\Delta x = -F'(x)F(x) \neq 0$. Then there exists $\mu > 0$ such that

$$T(x + s\Delta x^k) \; < \; T(x) \quad \text{for all} \quad 0 < s < \mu \; . \tag{1.246}$$

Proof. Define the function

$$\varphi(s) \; := \; T(x + s\Delta x) \; .$$

Obviously, $\varphi \in C^1(D_1)$ for some $D_1 \subset \mathbb{R}^1$ and

$$\varphi'(0) \; := \; (\text{grad } T(x + s\Delta x))^T|_{s=0} \; \Delta x \; =$$

$$= \; \underbrace{(F'(x)^T F(x))}_{= \; -\Delta x}{}^T \; \Delta x \; = \; - \; \|\Delta x\|^2 \; ,$$

which proves (1.246).

•

The **steplength strategy** consists of two parts:
a **reduction strategy** and a **prediction strategy**.
The **reduction strategy** applies when the monotonicity test fails, i.e.,

$$T(x^k + s_k^0 \delta x^k) \; > \; T(x^k)$$

for some s_k^0. In this case, the monotonicity test will be repeated with the reduced steplengths

$$s_k^{i+1} \; := \; \kappa \; s_k^i \quad , \quad i \in \mathbb{N}_0 \quad , \quad \kappa < 1 \quad (e.g., \; \kappa \approx \frac{1}{2} \; . \tag{1.247}$$

the downhill property (1.246) assures that a finite number of reductions will result in a **feasible steplength** $s_k^* > 0$.
The **prediction strategy** selects s_{k+1}^0 on the basis of an **ad-hoc rule** with respect to the **steplength history**

$$s_{k+1}^0 \; := \; \begin{cases} \min \; (s_{max}, \frac{s_k^*}{\kappa}) & , \quad \text{if } s_{k-1}^* \leq s_k^* \\ s_k^* & , \quad \text{otherwise} \end{cases} \quad . \tag{1.248}$$

Remark 5.1 The speed of convergence of the steepest descent method may be slow and problems may occur due to an ill-conditioning of the Jacobian $F'(x)$.

The steepest descent method as described above is not affine covariant. Indeed, given a nonsingular matrix $A \in \mathbb{R}^{n\times n}$, we may introduce another level function

$$T_A(x) \;:=\; \frac{1}{2}\,\|AF(x)\|^2 \;. \tag{1.249}$$

The following result underpins the necessity to develop an **affine covariant descent concept**, since it can be shown that almost always there exists a matrix A such that Δx is uphill with respect to T_A.

Lemma 5.2 Deficiency of the residual level function

Let $\Delta x = -\text{grad}\, T(x)$ be the descent direction with respect to the level function T . Then, unless

$$F'(x)\;\Delta x \;=\; \chi\; F(x) \quad \text{for some} \quad \chi < 0\;, \tag{1.250}$$

there exists a class of regular matrices A such that for some $\nu > 0$

$$T_A(x + s\Delta) \;>\; T_A(x) \quad , \quad 0 < s < \nu\;. \tag{1.251}$$

Proof. We have

$$\Delta x^T\;\text{grad}\; T_A(x) \;=\; -\;(F'(x)^T F(x))^T\;F'(x)^T A^T AF(x) \;=$$

$$=\; -\;F(x)^T\;F'(x)F'(x)^T A^T AF(x)\;.$$

We choose $A \in \mathbb{R}^{n\times n}$ such that

$$A^T A \;=\; F'(x)F'(x)^T \;+\; \mu\; yy^T \quad , \quad \mu > 0\;,$$

where $y \in \mathbb{R}^n$ satisfies

$$F(x)^T\;\Big(F'(x)F'(x)^T + I\Big)y \;=\; 0 \quad , \quad F(x)^T y \;\neq\; 0\;.$$

The existence of such an $y \in \mathbb{R}^n$ is guaranteed regarding that (1.250) is excluded.
We then get

$$\Delta x^T\;\text{grad}\; T_A(x) \;=$$

$$=\; -\;F(x)^T\;F'(x)F'(x)^T\Big(F'(x)F'(x)^T + \mu\; yy^T\Big)F(x) \;=$$

$$=\; -\;\|F'(x)F'(x)^T F(x)\|^2 \;+\; \mu F(x)^T\;yy^T F(x) \;=$$

$$= \; - \; \|F'(x)F'(x)^T F(x)\|^2 \; + \; \mu (F(x)^T y)^2 \; .$$

Hence, if we choose

$$\mu \; > \; \frac{\|F'(x)F'(x)^T F(x)\|^2}{(F(x)^T y)^2} \; ,$$

we get

$$\Delta x^T \; \text{grad} \; T_A(x) \; > \; 0 \; ,$$

which proves (1.251).

•

In order to come up with an **affine covariant globalization concept**, we introduce the level set associated with the level function T_A given by

$$G_A(z) \; := \; \{x \in D \mid T_A(x) \; \leq \; T_A(z)\} \; . \tag{1.252}$$

We recall that **monotonicity** with respect to T_A reads as follows

$$x^{k+1} \in \text{int} \; G_A(x^k) \quad , \quad \text{if} \quad \text{int} \; G_A(x^k) \neq \emptyset \; .$$

Denoting by $GL(n)$ the set of all regular $n \times n$ matrices, we introduce the **affine covariant level set**

$$\overline{G}_A(x) \; := \; \bigcap_{A \in GL(n)} G_A(x) \; . \tag{1.253}$$

Theorem 5.1 Newton path

Assume that $F : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$ is continuously differentiable on D with nonsingular Jacobi matrix $F'(x), x \in D$. Further suppose that for some $\hat{A} \in GL(n)$ the path-connected component of $G_{\hat{A}}(x^0), x^0 \in D$, is a compact subset of D. Then, the **path-connected component** of $\overline{G}_{\hat{A}}(x^0)$ is a **topological path** $\overline{x} : [0,2] \rightarrow \mathbb{R}^n$, called the **Newton path**. It has the properties

$$F(\overline{x}(\lambda)) \; = \; (1-\lambda) \; F(x^0) \; , \tag{1.254}$$

$$T_A(\overline{x}(\lambda)) \; = \; (1-\lambda)^2 \; T_A(x^0) \quad , \quad A \in GL(n) \; , \tag{1.255}$$

and satisfies the **two-point boundary value problem**

$$\begin{aligned} \frac{d\overline{x}}{d\lambda} \; &= \; - \; F'(\overline{x})^{-1} F(x^0) \; , \\ \overline{x}(0) \; &= \; x^0 \quad , \quad \overline{x}(1) \; = \; x^* \; . \end{aligned} \tag{1.256}$$

Moreover, we recover the **ordinary Newton increment** Δx^0 by means of

$$\frac{d\overline{x}}{d\lambda}|_{\lambda=0} \;=\; -\; F'(x^0)^{-1}F(x^0) \;=\; \Delta x^0 \;. \tag{1.257}$$

Proof. We introduce the level sets

$$H_A(x^0) \;:=\; \{y \in \mathbb{R}^n \mid \|Ay\|^2 \;\leq\; \|AF(x^0)\|^2\}$$

and define their intersection

$$\overline{H}(x^0) \;:=\; \bigcap_{A \in GL(n)} H_A(x^0) \;. \tag{1.258}$$

The idea of proof is to show that $\overline{H}(x^0) = \overline{G}(x^0)$.
For that purpose, we refer to $\sigma_i, 1 \leq i \leq n$, as the **singular values** of A and to $q_i, 1 \leq i \leq n$, as the associated **eigenvectors** of $A^T A$ such that

$$A^T A \;=\; \sum_{i=1}^{n} \sigma_i^2 \; q_i q_i^T \;.$$

We further denote by $\mathcal{A}$ the following subset of $GL(n)$

$$\mathcal{A} \;:=\; \{A \in GL(n) \mid A^T A \;=\; \sum_{i=1}^{n} \sigma_i^2 \; q_i q_i^T \;,\; q_1 \;=\; \frac{F(x^0)}{\|F(x^0)\|} \;\} \;.$$

Obviously, every $y \in \mathbb{R}^n$ admits the representation

$$y \;=\; \sum_{j=1}^{n} b_j q_j \quad , \quad b_j \in \mathbb{R} \;,\; 1 \leq j \leq n \;,$$

and hence,

$$\|Ay\|^2 \;=\; y^T A^T A y \;=\; \sum_{i=1}^{n} \sigma_i^2 b_i^2 \;,$$

$$\|AF(x^0)\|^2 \;=\; \sigma_1^2 \; \|F(x^0)\|^2 \;.$$

In particular, for $A \in \mathcal{A}$ we find

$$H_A(x^0) \;=\; \{y \in \mathbb{R}^n \mid \sum_{i=1}^{n} \sigma_i^2 b_i^2 \;\leq\; \sigma_1^2 \; \|F(x^0)\|^2\} \;.$$

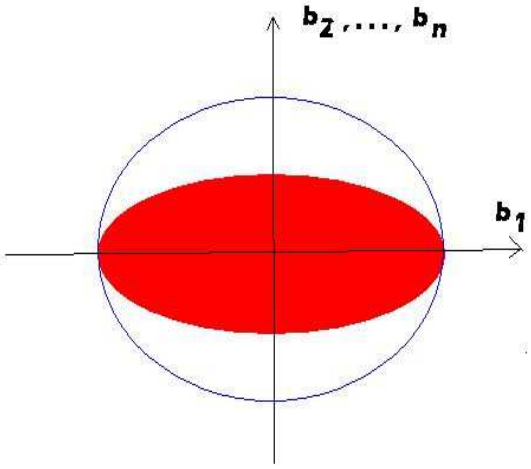

Figure 1: Intersection of ellipsoids $H_A(x^0), A \in \mathcal{A}$.

In other words, $H_A(x^0)$ defines the **n-dimensional ellipsoid**

$$\frac{1}{\|F(x^0)\|^2} b_1^2 \; + \; \left(\frac{\sigma_2}{\sigma_1 \|F(x^0)\|}\right)^2 b_2^2 \; + \; \ldots \; + \; \left(\frac{\sigma_n}{\sigma_1 \|F(x^0)\|}\right)^2 b_n^2 \; \leq \; 1 \; .$$

For $A \in \mathcal{A}$, all ellipsoids have a common b_1-axis of length $\|F(x^0)\|$, whereas the lengths of the other axes differ (cf. Figure 1).

It follows readily that

$$\hat{H}(x^0) \; = \; \{y \in \mathbb{R}^n \mid y \; = \; b_1 q_1 \; , \; |b_1| \leq \|F(x^0)\|\} \; = \tag{1.259}$$

$$= \; \{y \in \mathbb{R}^n \mid y = (1-\lambda) F(x^0) \; , \; \lambda \in [0,2]\} \; =$$

$$= \; \{y \in \mathbb{R}^n \mid Ay = (1-\lambda) AF(x^0) \; , \; \lambda \in [0,2] \; , \; A \in GL(n)\} \; .$$

Since $\mathcal{A} \subset GL(n)$, we have

$$\overline{H}(x^0) \subset \hat{H}(x^0) \; .$$

On the other hand, for $y \in \hat{H}(x^0)$ and $A \in \mathcal{A}$

$$\|Ay\|^2 \; = \; (1-\lambda)^2 \|AF(x^0)\|^2 \; \leq \; \|AF(x^0)\|^2 \; ,$$

which shows

$$\hat{H}(x^0) \subset \overline{H}(x^0) \; .$$

The final stage of the proof is done by an appropriate **lifting** of the path $\overline{H}(x^0)$ to $\overline{G}(x^0)$ using the **homotopy**

$$\Phi(x,\lambda) \;:=\; F(x) \;-\; (1-\lambda)F(x^0)\;.$$

In view of

$$\Phi_x \;=\; F'(x) \quad , \quad \Phi_\lambda \;=\; F(x^0)$$

and observing that Φ_x is nonsingular for $x \in D$ and $G_{\hat{A}}(x^0) \subset D$, **local continuation** from $\overline{x}(0) = x^0$ by the **implicit function theorem**, applied to $\Phi \equiv 0$, delivers the existence of the path

$$\overline{x} \subset G_{\hat{A}}(x^0) \subset D$$

with the properties (1.256),(1.257). The assertions (1.254) and (1.255) are now a direct consequence of (1.259).

•

Remark 5.2 The implication of the previous theorem is that even far from the solution, the Newton increment $\Delta x^0/\|\Delta x^0\|$, which is tangent to the Newton path originating from x^0, plays a decisive role and should be used in an affine invariant globalization strategy. Alone, its length may me too large and thus has to be controlled appropriately.

Remark 5.3 The previous theorem assumes that the Jacobian is regular in D. However, sometimes the situation is encountered where the Jacobian is singular at a **critical point** $\hat{x}$ even close to the initial guess x^0. In this case, the implicit function theorem tells us that the Newton path ends at that critical point.

5.2 Trust region concepts

As we have seen, far away from the solution the ordinary Newton method can be still used, provided an appropriate damping of the Newton increment is provided. Of course, we would like to know how to determine the damping factor, or in other words, what is the region around the current iterate where we can rely on the linearization with respect to the tangent to the Newton path. The specification of such regions is known as **trust region concepts**.

5.2.1 Trust region based on the Levenberg-Marquardt method

Given a current iterate $x^k \in \mathbb{R}^n$ and a prespecified parameter $\delta > 0$, the idea of the **Levenberg-Marquardt method** is to determine an increment $\Delta x^k \in \mathbb{R}^n$ as the solution of the **constrained minimization problem**

$$\inf_{\Delta x^k \in K_\delta} \|F(x^k) + F'(x^k)\Delta x^k\|\;,$$

where K_δ stands for the **constraint**

$$K_\delta \;:=\; \{\Delta x^k \in \mathbb{R}^n \mid \|\Delta x^k\| \leq \delta\} \;.$$

Coupling the inequality constraints by a **Lagrangian multiplier** $\mu \in \mathbb{R}_+$ leads to the **saddle point problem**

$$\inf_{\Delta x^k \in \mathbb{R}^n} \sup_{\mu \in \mathbb{R}_+} \mathcal{L}(\Delta x^k, \mu)$$

in terms of the associated **Lagrangian functional**

$$\mathcal{L}(\Delta x^k, \mu) \;:=\; \|F(x^k) + F'(x^k)\Delta x^k\|^2 \;+\; \mu \left(\|\Delta x^k\|^2 \;-\; \delta^2\right) .$$

The **KKT conditions** read as follows:

$$\left(F'(x^k)^T F'(x^k) + \mu I\right)\Delta x^k \;=\; -\; F'(x^k)F(x^k) \;, \tag{1.260}$$

$$\mu \geq 0 \quad , \quad \|\Delta x^k\|^2 - \delta^2 \leq 0 \quad , \quad \mu\; (\|\Delta x^k\|^2 - \delta^2) \;=\; 0 \;. \tag{1.261}$$

Denoting the solution of the saddle point problem by $(\Delta x^k(\mu), \mu)$, we observe

$$\mu \to 0_+ \quad \Longrightarrow \quad \Delta x^k(\mu) \to -\; F'(x^k)F(x^k) \;,$$

$$\mu >> 1 \quad \Longrightarrow \quad \Delta x^k(\mu) \approx -\; \frac{1}{\mu}\; F'(x^k)F(x^k) \;=\; -\; \frac{1}{\mu}\; \text{grad}\; T(x^k) \;.$$

This means:
Close to the solution, the method coincides with the ordinary Newton method, whereas far from the solution, it corresponds to a steepest descent with the steplength parameter $\frac{1}{\mu}$.

The Levenberg-Marquardt method looks robust, since the coefficient matrix $F'(x^k)^T F'(x^k) + \mu I$ in (1.260) is regular, even if the Jacobian $F'(x^k)$ is singular. However, the method may terminate for singular $F'(x^k)$, since then the right-hand side in (1.260) also degenerates. Moreover, the Levenberg-Marquardt methods lacks affine invariance.

5.2.2 The Armijo damping strategy

An empirical damping strategy is the **Armijo strategy**:
Let $\Lambda_k \subset \{1, \frac{1}{2}, \frac{1}{4}, ..., \lambda_{min}\}$ be a sequence of steplengths with the property

$$T(x^k \;+\; \lambda \Delta x^k) \;\leq\; (1 - \frac{1}{2}\lambda)\; T(x^k) \quad , \quad \lambda \in \Lambda_k \;. \tag{1.262}$$

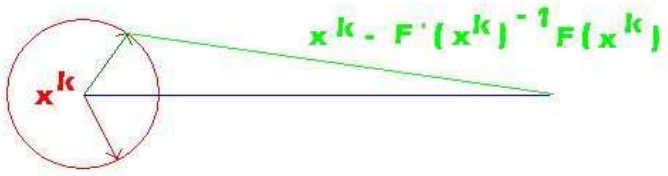

Figure 2: Geometric interpretation of the affine covariant trust region method

Then, the **damping parameter** $\lambda_k \in \Lambda_k$ is chosen as the optimal one:

$$T(x^k + \lambda_k \Delta x^k) = \min_{\lambda \in \Lambda_k} T(x^k + \lambda \Delta x^k) \ .$$

Obviously, the choice of the level function $T(x)$ in the Armijo rule does not reflect affine covariance. We will develop an affine covariant damping strategy below.

5.2.3 Affine covariant trust region method

The Levenberg-Marquardt method can be easily reformulated to yield an affine covariant version. Since affine covariance means affine invariance with respect to transformations in the domain of definition, we have to modify the objective functional:

$$\inf_{\Delta x^k \in K_\delta} \| F'(x^k)^{-1} \Big(F(x^k) + F'(x^k) \Delta x^k \Big) \| \ , \tag{1.263}$$

whereas the set of constraints K_δ is given as before.

The affine covariant trust region method (1.263) admits an easy geometric interpretation as shown in Figure 5.2. The set K_δ of constraints is represented as a sphere with radius δ around x^k. If δ exceeds the length of the Newton correction Δx^k, the constraint is not active, and we are in the regime of the ordinary Newton method. However, if δ is smaller than the Newton correction Δx^k, we have to apply an appropriate damping.

5.2.4 Affine contravariant trust region method

We can also easily reformulate the Levenberg-Marquardt method to come up with an affine contravariant version. Since affine contravariance means affine invariance with respect to transformations in the range space, the objective functional remains unchanged, but we have to modify the set of constraints:

$$\inf_{\Delta x^k \in \tilde{K}_\delta} \|F(x^k) + F'(x^k)\Delta x^k\| \ ,$$

whereas the set of constraints $\tilde{K}_\delta$ is given as follows:

$$\tilde{K}_\delta \ := \ \{\Delta x^k \in \mathbb{R}^n \mid \|F'(x^k)\Delta x^k\| \ \leq \ \delta\} \ . \tag{1.264}$$

There is basically the same geometric interpretation as before with the only difference that now the picture has to be drawn in the range space.

5.3 Globalization of affine contravariant Newton methods

5.3.1 Convergence of the damped Newton iteration

We consider the **damped Newton iteration**

$$\begin{aligned} F'(x^k)\Delta x^k \ &= \ -\ F(x^k) \ , \\ x^{k+1} \ &= \ x^k \ + \ \lambda_k \Delta x^k \ , \ \lambda_k \in [0,1] \end{aligned} \tag{1.265}$$

in an affine contravariant setting where the damping factor λ_k is chosen to achieve residual contraction.

Theorem 5.2 Optimal choice of the damping factor

Assume that $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$, D convex, is continuously differentiable on D with regular Jacobian $F'(x), x \in D$. We further suppose that the following **affine contravariant Lipschitz condition** holds true

$$\|\Big(F'(y) - F'(x)\Big)(y-x)\| \ \leq \ \omega \ \|F'(x)(y-x)\|^2 \quad , \quad x, y \in D \ . \tag{1.266}$$

Setting $h_k := \omega \|F(x^k)\|$, for $\lambda \in [0, \min(1, \frac{2}{h_k})]$ we have

$$\|F(x^k + \lambda \Delta x^k)\| \ \leq \ t_k(\lambda) \ \|F(x^k)\| \ , \tag{1.267}$$

where

$$t_k(\lambda) \ := \ 1 \ - \ \lambda \ + \ \frac{1}{2} \ h_k \ \lambda^2 \ .$$

The **optimal choice of the damping factor** is

$$\lambda_k^* \;:=\; \min(1, \frac{1}{h_k}) \;. \tag{1.268}$$

Proof. By straightforward calculation we find

$$\|F(x^k + \lambda\Delta x^k)\| \;=\; \|F(x^k + \lambda\Delta x^k \;-\; F(x^k) \;-\; F'(x^k)\Delta x^k\| \;=$$

$$=\; \| \int_0^\lambda \Big(F'(x^k + t\Delta x^k) - F'(x^k) \Big)\; \Delta x^k \; dt \;-\; (1-\lambda)\; F'(x^k)\Delta x^k\| \;\leq$$

$$\leq\; \| \int_0^\lambda \Big(F'(x^k + t\Delta x^k) - F'(x^k) \Big)\; \Delta x^k \; dt\| \;+\; (1-\lambda)\; \|F(x^k)\| \;.$$

The first term on the right-hand side measures the **deviation from the Newton path**. Using the affine contravariant Lipschitz condition, it can be estimated as follows

$$\| \int_0^\lambda \Big(F'(x^k + t\Delta x^k) - F'(x^k) \Big)\; \Delta x^k \; dt\| \;\leq$$

$$\leq\; \frac{1}{2}\omega\lambda^2\; \|F'(x^k)\Delta x^k\|^2 \;\leq\; \frac{1}{2}h_k\lambda^2\; \|F(x^k)\| \;.$$

Inserting this estimate into the previous one and minimizing $t_k(\lambda)$ proves the theorem.

•

Theorem 5.3 Global convergence of affine contravariant Newton methods

Under the same assumptions as in theorem 5.2 let D_0 be the path-connected component of the level set $G(x^0)$ and suppose that D_0 is a compact subset of D. Then, the for all damping factors

$$\lambda_k \in [\varepsilon, 2\lambda_k^* - \varepsilon] \tag{1.269}$$

with $\varepsilon > 0$ sufficiently small, the damped Newton iterates $x^k, k \in \mathbb{N}_0$ converge to some $x^* \in D_0$ with $F(x^*) = 0$.

Proof. The parabola $t_k(\lambda)$ from Theorem 5.2 can be bounded by a polygonal as follows

$$t_k(\lambda) \;\leq\; \begin{cases} 1 \;-\; \frac{1}{2}\lambda & , \quad 0 \leq \lambda \leq \frac{1}{h_k} \;, \\ 1 \;+\; \frac{1}{2}\lambda \;-\; \frac{1}{h_k} & , \quad \frac{1}{h_k} \leq \lambda \leq \frac{2}{h_k} \;. \end{cases}$$

For $0 < \varepsilon \leq \frac{1}{h_k}$ and $\lambda_k \in [\varepsilon, 2\lambda_k^* - \varepsilon]$ we thus have

$$t_k(\lambda) \;\leq\; 1 \;-\; \frac{1}{2}\varepsilon \;, \tag{1.270}$$

which shows strict reduction of the residual level function $T(x)$.
The existence of a global $\varepsilon > 0$ follows from the compactness assumption on D_0 which implies

$$\max_{x \in D_0} \|F(x)\| \;<\; \infty \;.$$

Consequently, if $G(x^k) \subset D_0$, then (1.270) yields

$$G(x^{k+1}(\lambda)) \subset G(x^k) \;.$$

The rest of the proof is along the same lines as the proof of the affine contravariant Newton-Mysovskikh theorem.

•

5.3.2 Adaptive affine contravariant trust region strategy

In Theorem 5.2 we derived the theoretical damping factor (1.268). Since the Kantorovich quantity $h_k = \omega\|F(x^k)\|$ cannot be accessed directly, we again have to provide appropriate estimates

$$[h_k] \;:=\; [\omega] \;\|F(x^k)\| \;, \tag{1.271}$$

where $[\omega]$ is a lower bound for the domain dependent Lipschitz constant that can be obtained by **pointwise sampling**.
Then, an estimate of the **optimal damping factor** is given by means of

$$[\lambda_k^*] \;:=\; \min\,(1, \frac{1}{[h_k]}) \;. \tag{1.272}$$

It follows readily from (1.271) that

$$[\lambda_k^*] \;\geq\; \lambda_k^* \;,$$

i.e., we may have a considerable **overestimation**. As a remedy, repeated reductions must be performed by appropriate **prediction and correction strategies**.
The following **bit counting lemma** gives information about the contraction in the residuals in terms of the accuracy of the estimate for the Kantorovich quantity.

Lemma 5.3 Bit counting lemma
Assume that for some $0 \leq \sigma < 1$ there holds

$$0 \;\leq\; h_k \;-\; [h_k] \;<\; \sigma \max\,(1, [h_k])\;. \tag{1.273}$$

Then, the **residual monotonicity test** (1.267) yields

$$\|F(x^{k+1})\| \;\leq\; \left(1 - \frac{1}{2}(1-\sigma)\lambda_k^*\right)\,\|F(x^k)\|\;. \tag{1.274}$$

Proof. The assumption (1.273) can be rewritten as

$$[h_k] \;\leq\; h_k \;<\; (1+\sigma)\,\max\,(1, [h_k])\;,$$

which results in the following estimate of the residual contraction

$$\frac{\|F(x^{k+1})\|}{\|F(x^k)\|} \;\leq\; [1 - \lambda + \frac{1}{2}\,\lambda^2 h_k]|_{\lambda=[\lambda_k^*]} \;<$$

$$<\; [1 - \lambda + \frac{1}{2}\,(1+\sigma)\,\lambda^2 [h_k]]|_{\lambda=[\lambda_k^*]} \;\leq\; 1 - \frac{1}{2}\,(1-\sigma)\,\lambda_k^*\;.$$

•

Remark 5.4 Restricted residual monotonicity test
For $\sigma \leq \frac{1}{2}$, (1.274) results in the **restricted residual monotonicity test**

$$\|F(x^{k+1})\| \;\leq\; \left(1 - \frac{1}{4}\lambda_k^*\right)\,\|F(x^k)\|\;. \tag{1.275}$$

This should be compared with the heuristically derived **Armijo strategy** (1.262).

In order to come up with a **computationally feasible, affine contravariant adaptive trust region strategy**, we have to provide appropriate estimates of the Kantorovich quantities.
For the k-th iteration step, such a strategy consists of a **correction step**, providing a reliable damping factor, and a **prediction step**, predicting an initial guess λ_{k+1}^0 for the subsequent $k+1$-st iteration step.

As far as the **correction step** is concerned, we recall that the damped Newton method with damping factor $\lambda \in [0,1]$ represents a **deviation from the Newton path** which can be measured by means of

$$\|F(x^{k+1}) - (1-\lambda)F(x^k)\| \;\leq\; \frac{1}{2}\,\omega\,\lambda^2\,\|F(x^k)\|^2 \;.$$

This leads us to the following **lower bound** for the affine contravariant Kantorovich quantity

$$[h_k] \;:=\; \frac{2\,\|F(x^{k+1}) - (1-\lambda)F(x^k)\|}{\lambda^2\,\|F(x^k)\|} \;\leq\; h_k \;.$$

Using the prediction λ_k^0 from the previous step, for $i \geq 0$ we compute the **trial iterate**

$$x^{k+1} \;=\; x^k \;+\; \lambda_k^i \Delta x^k$$

and perform the **residual monotonicity test**

$$\|F(x^{k+1})\| \;\leq\; (1 \;-\; \frac{1}{4}\,\lambda_k^i)\,\|F(x^k)\| \;.$$

If the test is successful, we accept the current value λ_k^i as the damping factor. Otherwise, we set

$$\lambda_k^{i+1} \;=\; \min\,(\frac{1}{2}\lambda_k^i, \frac{1}{[h_k^i]}) \;.$$

As long as $\lambda_k^{i+1} \geq \lambda_{min}$, the gives us a new trial iterate. However, if $\lambda_k^{i+1} < \lambda_{min}$, the process is stopped (**convergence failure**).
For the **prediction** of a damping factor λ_{k+1}^0, we recall

$$h_{k+1} \;=\; \frac{\|F(x^{k+1})\|}{\|F(x^k)\|}\;h_k \;.$$

Denoting by i_* the index, for which $\lambda_k^{i_*}$ passed the residual monotonicity test, we use the lower bound

$$[h_{k+1}^0] \;=\; \frac{\|F(x^{k+1})\|}{\|F(x^k)\|}\;[h_k^{i_*}] \;<\; [h_k^{i_*}]$$

and set

$$\lambda_{k+1}^0 \;:=\; \min\,(1, \frac{1}{[h_{k+1}^0]}) \;.$$

Obviously, we need an initial guess λ_0^0 which is chosen as $\lambda_0^0 = 1$ for mildly nonlinear problems and $\lambda_0^0 \ll 1$ for highly nonlinear problems.

5.4 Error oriented descent

We consider the **damped Newton iteration**

$$\begin{aligned} F'(x^k)\Delta x^k &= -F(x^k)\ , \\ x^{k+1} &= x^k + \lambda_k \Delta x^k\ ,\ \lambda_k \in [0,1] \end{aligned} \tag{1.276}$$

in an affine covariant framework. Consequently, the damping factor λ_k has to be chosen in such a way that the deviation from the Newton path is controlled in lights of an affine covariant Lipschitz condition. This will lead to the **natural monotonicity test**

$$\|\overline{\Delta x}^{k+1}\| < \|\Delta x^k\|\ , \tag{1.277}$$

where $\overline{\Delta x}^{k+1}$ stands for the **simplified Newton correction**

$$F'(x^k)\overline{\Delta x}^{k+1} = -F(x^{k+1})\ . \tag{1.278}$$

5.4.1 General level functions

We start from the following, easily verifiable **local descent result** for the Newton direction with respect to general level functions

$$T_A(x) := \frac{1}{2}\|AF(x)\|^2 \quad , \quad A \in \mathrm{GL}(n)\ . \tag{1.279}$$

Lemma 5.4 General downhill property

Assume that $F : D \subset \mathbb{R}^n \to \mathbb{R}^n, D \subset \mathbb{R}^n$ convex, is continuously differentiable. Then, for all $A \in \mathrm{GL}(n)$ there holds

$$(\Delta x^k)^T \ \mathrm{grad}\ T_A(x) = -2\,T_A(x) < 0\ . \tag{1.280}$$

The previous result tells us that with regard to **first order information** all level functions are equally well suited. In order to be more selective, we have to use **second order information**.

Theorem 5.4 Affine covariant downhill property

Assume that $F : D \subset \mathbb{R}^n \to \mathbb{R}^n, D \subset \mathbb{R}^n$ convex, is continuously differentiable on D with regular Jacobian $F'(x), x \in D$ and suppose further that the following **affine covariant Lipschitz condition** holds true

$$\|F'(x)^{-1}\Big(F'(y) - F'(x)\Big)(y-x)\| \le \omega\,\|y-x\|^2 \quad , \quad x, y \in D\ . \tag{1.281}$$

Let $x^k \in D$ be an iterate such that

$$G_A(x^k) \subset D \quad , \quad A \in \mathrm{GL}(n) \; , \tag{1.282}$$

and denote by h_k and $\overline{h}_k$ the **Kantorovich quantities**

$$h_k \;:=\; \omega \, \|\Delta x^k\| \quad , \quad \overline{h}_k \;:=\; h_k \, \mathrm{cond}(AF'(x^k)) \; . \tag{1.283}$$

Then, for all $\lambda \in [0, \min(1, 2/\overline{h}_k)]$ there holds

$$\|AF(x^k + \lambda \Delta x^k)\| \;\leq\; t_k^A(\lambda) \, \|AF(x^k)\| \; , \tag{1.284}$$

where

$$t_k^A(\lambda) \;:=\; 1 \;-\; \lambda \;+\; \frac{1}{2} \, \lambda^2 \, \overline{h}_k \; . \tag{1.285}$$

The **optimal damping factor** is given by

$$\lambda_k^*(A) \;:=\; \min \, (1, 1/\overline{h}_k) \; . \tag{1.286}$$

Proof. For $\lambda \in [0, 1]$, we have

$$\|AF(x^k + \lambda \Delta x^k)\| \;=\; \|AF(x^k + \lambda \Delta x^k) - \; F(x^k) \; + \underbrace{F(x^k)}_{= \, -F'(x^k) \Delta x^k} \; \| \;=$$

$$= \; \|A \, \Big(\int_0^\lambda (F'(x^k + t \Delta x^k) - F'(x^k)) \; \Delta x^k \; dt \;-\; (1 - \lambda) \; \underbrace{F'(x^k) \Delta x^k}_{= \, -F(x^k)} \Big) \| \;\leq$$

$$\leq \; \|A \int_0^\lambda (F'(x^k + t \Delta x^k) - F'(x^k)) \; \Delta x^k \; dt \| \;+\; (1 - \lambda) \, \|AF(x^k)\| \; .$$

Invoking the affine covariant Lipschitz condition, for the first term on the right-hand side we obtain

$$\|AF'(x^k) \int_0^\lambda F'(x^k)^{-1} (F'(x^k + t \Delta x^k) - F'(x^k)) \; \Delta x^k \; dt \| \;\leq$$

$$\leq \; \|AF'(x^k)\| \int_0^\lambda \omega \; t \; \underbrace{t}_{\leq \, 1} \; \|\Delta x^k\| \; \| \underbrace{\Delta x^k}_{= \, -(AF'(x^k))^{-1} AF(x^k)} \| \; dt \leq$$

$$\leq \frac{1}{2}\,\lambda^2\,\|AF'(x^k)\|\;h_k\;\|(AF'(x^k))^{-1}AF(x^k)\|\;\leq$$

$$\leq \frac{1}{2}\,\lambda^2\,h_k\,\|AF'(x^k)\|\,\|(AF'(x^k))^{-1}\|\,\|AF(x^k)\|\;=\;\frac{1}{2}\,\lambda^2\,\overline{h}_k\,\|AF(x^k)\|\;.$$

Combining the previous estimates gives the assertions.

•

In view of Theorem 5.4 we readily get the following **global convergence result**.

Theorem 5.5 Affine covariant global convergence theorem

In addition to the assumptions of Theorem 5.4, let $x^0 \in D$ be an initial guess such that the **path-connected component** D_0 of $G_A(x^0)$ is a **compact subset** of D.
Then, for all damping parameters

$$\lambda_k \in [\varepsilon\;,\;2\;\lambda_k^*(A) - \varepsilon]\;, \tag{1.287}$$

with $\varepsilon > 0$ being sufficiently small, the **damped Newton method converges** to some $x^* \in D_0$ with $F(x^*) = 0$.

Proof. As before, we remark that the parabola $t_k^A(\lambda)$ can be bounded from above by a polygonal bound according to

$$t_k^A(\lambda)\;\leq\;1\;-\;\frac{1}{2}\,\varepsilon\quad,\quad 0 < \varepsilon \leq \frac{1}{\overline{h}_k}\;. \tag{1.288}$$

Moreover, there is a **global** ε, since with regard to the compactness assumption on D_0 we have

$$\max_{x\in D_0}\|F'(x)^{-1}F(x)\|\;\mathrm{cond}(AF'(x))\;<\;\infty\;.$$

The proof proceeds by induction on k: Assuming $G_A(x^k) \subseteq D_0$, (1.288) yields

$$G_A(x^{k+1}) \subset G_A(x^k) \subseteq D_0\;.$$

Consequently, the sequence of Newton iterates lives in a compact set which allows to conclude.

•

Remark 5.5 The flaws of residual monotonicity

Setting $A = I$ in the previous theorem, we are obviously back in the residual based regime where we have proved global convergence according to Theorem 5.3. However, if the Jacobian $F'(x^k)$ is **ill conditioned**, we obtain

$$\lambda_k^*\;=\;\Big(h_k\;\mathrm{cond}(F'(x^k))\Big)^{-1}\;\ll\;1\;, \tag{1.289}$$

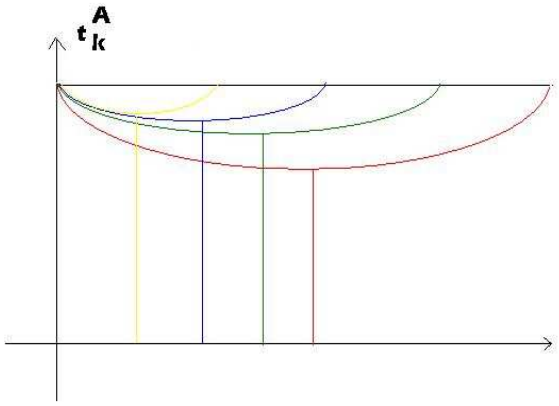

Figure 3: Reduction factors and optimal damping factors

which algorithmically will result in a **termination of the iteration**.

5.4.2 Natural level function

In view of (1.283) and (1.286), the **most natural choice** of the matrix $A \in \mathrm{GL}(n)$ in the level function T_A is

$$A \;:=\; A_k \;=\; F'(x^k)^{-1} \;. \tag{1.290}$$

The associated level function $T_{F'(x^k)^{-1}}$ is called the **natural level function** which gives rise to the **natural monotonicity test**

$$\|\overline{\Delta x}^{k+1}\| \;\leq\; \|\Delta x^k\| \tag{1.291}$$

in terms of the **simplified Newton correction**

$$\overline{\Delta x}^{k+1} \;=\; -\; F'(x^k)^{-1} F(x^{k+1}) \;. \tag{1.292}$$

Several remarks are due with respect to the **properties of the natural level function**.

Remark 5.6 Extremal properties

As shown in Figure 3, for $A \in \mathrm{GL}(n)$ the reduction factors $t_k^A(\lambda)$ and the optimal damping factors $\lambda_k^*(A)$ satisfy

$$t_k^{A_k}(\lambda) \;=\; 1 \;-\; \lambda \;+\; \frac{1}{2}\,\lambda^2\, h_k \;\leq\; t_k^A(\lambda) \;, \tag{1.293}$$

$$\lambda_k^*(A_k) \;=\; \min\,(1, \frac{1}{h_k}) \;\geq\; \lambda_k^*(A) \;. \tag{1.294}$$

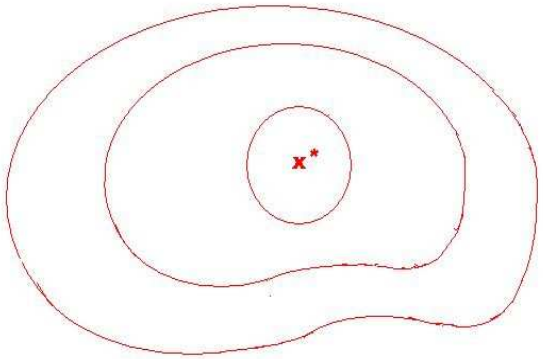

Figure 4: Asymptotic distance spheres associated with natural level sets

Remark 5.7 Steepest descent property

The damped Newton method in x^k is a **method of steepest descent** for the natural level function T_{A_k}:

$$\Delta x^k \;=\; -\,\mathrm{grad}\; T_{A_k}(x^k)\;. \tag{1.295}$$

Remark 5.8 Asymptotic optimality

In view of

$$h_k < 1 \quad\Longrightarrow\quad \lambda_k^*(A_k) = 1\;, \tag{1.296}$$

the damped Newton method asymptotically achieves **quadratic convergence**.

Remark 5.9 Asymptotic distance function

If $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$ is twice continuously differentiable, we can show

$$T_{F'(x^*)^{-1}}(x) \;=\; \frac{1}{2}\,\|x - x^*\|^2 \;+\; O(\|x - x^*\|^3)\;.$$

Hence, for $x^k \to x^*$ the natural monotonicity criterion approaches a **distance criterion** of the form

$$\|x^{k+1} - x^*\| \;\le\; \|x^k - x^*\|\;.$$

As shown in Figure 4, close to the solution x^* the natural level surface is close to a sphere, whereas it degenerates to an **osculating sphere** with increasing

distance to x^*. Note that for other level functions, the level surface is an ellipsoid close to x^*, with the ratio of the largest to the smallest half-axis being related to the condition number of the Jacobian, and an osculating ellipsoid off x^*.

Remark 5.10 Local descent

if we insert $A = A_k$ into (1.285),(1.286) of Theorem 5.4, we get the **local descent property**

$$\|\overline{\Delta x}^{k+1}\| \;\leq\; \Big(1 - \lambda + \frac{1}{2}\,\lambda^2\,h_k\Big)\,\|\Delta x^k\|\;. \tag{1.297}$$

Remark 5.11 Global convergence

We note that the results of Theorem 5.5 are not applicable to the situation at hand, since $A = A_k$ changes from one step to the other. Taking the asymptotic distance function property into account, in the subsequent **global convergence result** we make the fixed choice $A = F'(x^*)^{-1}$.

Theorem 5.6 Global convergence of the affine covariant damped Newton method with natural level functions; Part I

Assume that $F : D \subset \mathbb{R}^n \to \mathbb{R}^n, D \subset \mathbb{R}^n$ convex, is continuously differentiable on D with regular Jacobian $F'(x), x \in D$ and suppose that the following **affine covariant Lipschitz condition** is fulfilled

$$\|F'(x^*)^{-1}\Big(F'(y) - F'(x)\Big)(y - x)\| \;\leq\; \omega_*\,\|y - x\|^2 \quad,\quad x, y \in D\;. \tag{1.298}$$

Suppose further that $x^* \in D$ is the unique solution in D and let $x^0 \in D$ be an initial guess such that the path-connected component of $G_{F'(x^*)^{-1}}(x^0)$ is a compact subset of D.
Let the **damping factors** be chosen according to

$$\lambda_k \in [\varepsilon, 2\lambda_k^* - \varepsilon] \quad,\quad 0 < \varepsilon < \frac{1}{h_k^*}\;, \tag{1.299}$$

where

$$\lambda_k^* \;:=\; \min(1, \frac{1}{h_k^*}) \quad,\quad h_k^* \;:=\; \omega_*\,\|\Delta x^k\|\,\|F'(x^k)^{-1}F'(x^*)\|\;. \tag{1.300}$$

Then, the damped Newton iteration converges globally to x^*.

Proof. In the proof of Theorem 5.4 we have shown

$$\begin{aligned}
&\|A\;F(x^k + \lambda\Delta x^k)\| \;\leq \\
\leq\;& \|A\int_0^\lambda (F'(x^k + t\Delta x^k) - F'(x^k))\;\Delta x^k\;dt\| \;+\; (1-\lambda)\;\|AF(x^k)\|\;.
\end{aligned}$$

Choosing $A = F'(x^*)^{-1}$ and observing

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

the first term on the right-hand side of the previous inequality is now estimated as follows

$$\|F'(x^*)^{-1}\int_0^\lambda (F'(x^k+t\Delta x^k)-F'(x^k))\;\Delta x^k\;dt\| \;\leq$$

$$\frac{1}{2}\;\lambda^2\;\underbrace{\omega_*\;\|\Delta x^k\|\;\|F'(x^k)^{-1}F'(x^*)\|}_{=\,h_k^*}\;\|F'(x^*)^{-1}F(x^k)\|\;.$$

The rest of the proof proceeds in exactly the same manner as in the proof of Theorem 5.5.

•

In much the same way as we derived Theorem 5.5 from Theorem 5.4, the previous results imply the following convergence statement in a more realistic scenario:

Corollary 5.7 Global convergence of the affine invariant damped Newton method; Part II

Assume that all assumptions of Theorem 5.6 are met, except that the **affine covariant Lipschitz condition** is replaced by one with a **local Lipschitz constant**

$$\|F'(z)^{-1}\Big(F'(y)-F'(x)\Big)(y-x)\| \;\leq\; \omega(z)\;\|y-x\|^2 \quad,\quad x,y,z\in D_0 \tag{1.301}$$

Then, the damped Newton method converges for

$$\lambda\in[\varepsilon,2\lambda_k^*(z)-\varepsilon] \quad,\quad 0<\varepsilon<\frac{1}{h_k(z)}\;, \tag{1.302}$$

with the **optimal damping factor** given by

$$\lambda_k^*(z) \;:=\; \min\,(1,\frac{1}{h_k(z)})\;, \tag{1.303}$$

where

$$h_k(z) \;:=\; \omega(z)\;\|\Delta x^k\|\;\|F'(x^k)^{-1}F'(z)\|\;. \tag{1.304}$$

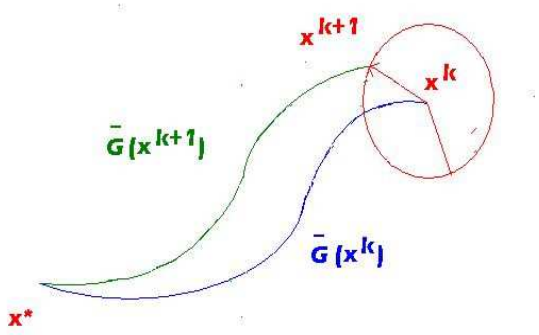

Figure 5: Newton path $\overline{G}(x^k)$, trust region around x^k and Newton step with locally optimal damping factor

We have a **local level function reduction** according to

$$T_{F'(z)^{-1}}(x^k + \lambda \Delta x^k) \;\leq\; \left(1 - \lambda + \frac{1}{2}\,\lambda^2\, h_k(z)\right)^2 T_{F'(z)^{-1}}(x^k)\;. \qquad (1.305)$$

Remark 5.12 Recovery of the exact Newton method

If we choose $z := x^k$, we recover the theoretically optimal damping strategy for the exact Newton method.

Remark 5.13 Geometrical interpretation

As shown in Figure 5, the damped Newton method proceeds along the tangent of the Newton path $\overline{G}(x^k)$ with the actual steplength

$$\|x^{k+1} - x^k\| \;=\; \lambda_k^*\, \|\Delta x^k\| \;\leq\; \rho_k \;:=\; \frac{1}{\omega_k}\;,$$

where the radius ρ_k describes the **local trust region** around the current iterate x^k.

5.4.3 Adaptive trust region strategies

We provide lower estimates

$$[\omega_k] \;\leq\; \omega_k \quad , \quad [h_k] \;\leq\; h_k \qquad (1.306)$$

for the Lipschitz constant and the Kantorovich quantity (e.g., by pointwise sampling of the domain, and thus get an **upper estimate**

$$[\overline{\lambda}_k] \;:=\; \min(1, \frac{1}{[h_k]}) \;\geq\; \overline{\lambda}_k \qquad (1.307)$$

of the **damping factor**. Since (1.307) usually leads to an overestimation, we have to perform an appropriate **prediction** and **correction strategies** which depend on the required accuracy.

Lemma 5.5 Bit counting lemma

Assume that for some $0 \leq \sigma < 1$ there holds

$$0 \;\leq\; h_k - [h_k] \;\leq\; \sigma \max(1, [h_k]) \;. \tag{1.308}$$

then, the **natural monotonicity test** gives

$$\|\overline{\Delta x}^{k+1}\| \;\leq\; (1 - \frac{1}{2}(1-\sigma)\lambda)\|\Delta x^k\| \;. \tag{1.309}$$

Proof. The proof is left as an exercise. •

Remark 5.14 Restricted natural monotonicity test

For $\sigma \leq \frac{1}{2}$, the bit counting lemma suggests the following **restricted natural monotonicity test**

$$\|\overline{\Delta} x^{k+1}\| \;\leq\; \Big(1 - \frac{\lambda}{4}\Big) \;\|\Delta x^k\| \;. \tag{1.310}$$

Correction strategy

We have to monitor the deviation from the Newton path. In an **affine covariant setting** we have the upper bound

$$\|\overline{\Delta} x^{k+1}(\lambda) - (1-\lambda)\;\Delta x^k\| \;\leq\; \frac{1}{2}\;\lambda^2\;\omega_k\;\|\Delta x^k\|^2 \;,$$

which gives us the following estimate for the **Kantorovich quantity**

$$[h_k](\lambda) \;=\; [\omega_k]\;\|\Delta x^k\| \;:=\; \frac{2\|\overline{\Delta} x^{k+1}(\lambda) - (1-\lambda)\;\Delta x^k\|}{\lambda^2\;\|\Delta x^k\|} \;\leq\; h_k \;. \tag{1.311}$$

Assuming that we have a **trial value** λ_k^0 at disposal, for $j \geq 0$ we compute

$$\lambda_k^{j+1} \;:=\; \min(\frac{\lambda_k^j}{2}, [h_k](\lambda_k^j)) \tag{1.312}$$

as long as the restricted natural monotonicity test fails.

Prediction strategy

The prediction strategy aims to provide a reasonable initial estimate λ_k^0. Such an estimate can only be accessed, if we use a **Lipschitz constant** $\overline{\omega}_k$ satisfying

$$\|F'(x^k)^{-1}\Big(F'(x) - F'(x^k)\Big)v\| \leq\; \overline{\omega}_k\;\|x - x^k\|\;\|v\| \;,$$

where v is supposed in some sense "close" to $x - x^k$.
We are thus led to the local estimate

$$\begin{aligned} \|\overline{\Delta}x^k - \Delta x^k\| &= \|\Big(F'(x^{k-1})^{-1} - F'(x^k)^{-1}\Big)F(x^k)\| = \\ &= \|F'(x^k)^{-1}\Big(F'(x^k) - F'(x^{k-1})\Big)\overline{\Delta}x^k\| \leq \overline{\omega}_k \, \lambda_{k-1} \, \|\Delta x^{k-1}\| \, \|\overline{\Delta}x^k\| \, , \end{aligned}$$

which suggests the estimate

$$[\overline{\omega}_k] := \frac{\|\overline{\Delta}x^k - \Delta x^k\|}{\lambda_{k-1} \, \|\Delta x^{k-1}\| \, \|\overline{\Delta}x^k\|} \leq \overline{\omega}_k \, . \tag{1.313}$$

The estimate (1.313) exploits "newest information" and leads us to the **prediction strategy**

$$\lambda_k^0 := \min(1, \mu_k) \quad , \quad \mu_k := \frac{\|\Delta x^{k-1}\|}{\|\overline{\Delta}x^k - \Delta x^k\|} \frac{\|\overline{\Delta}x^k\|}{\|\Delta x^k\|} \lambda_{k-1} \, . \tag{1.314}$$

Finally, as far as an initial guess λ_0^0 is concerned, we choose $\lambda_0^0 = 1$ for mildly nonlinear problems and $\lambda_0^0 \ll 1$ for highly nonlinear problems.

6. Continuation Methods for Parameter Dependent Systems

In applications, frequently nonlinear systems occur that depend on one or several parameters representing the influence of specific quantities on the behavior of the systems. A classical example for such a **parameter dependent nonlinear system** is the **Bratu problem**

$$\begin{aligned} -\Delta u &= \lambda \exp(u) \quad \text{in } \Omega \subset \mathbb{R}^3 \,, \\ u &= 0 \qquad\quad\;\; \text{on } \Gamma = \partial\Omega \,, \end{aligned} \tag{1.315}$$

describing certain exothermal chemical reactions, where $\lambda > 0$ stands for the so-called Arrhenius parameter.

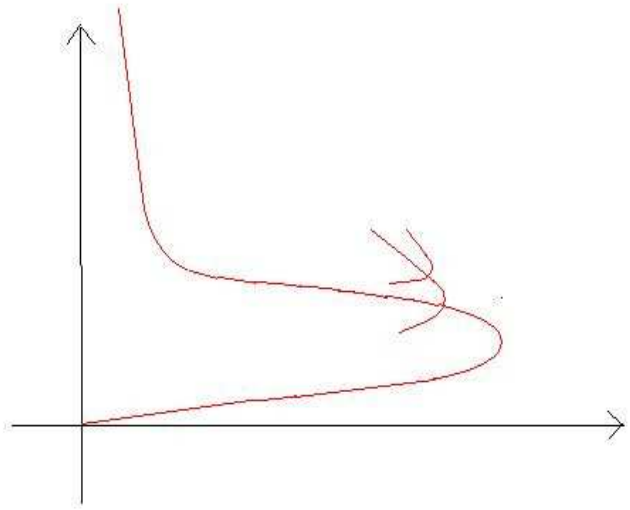

Figure 6: Bifurcation diagram for the Bratu problem

An important feature of such problems is that they typically exhibit **multiple solutions** and **bifurcation phenomena**. Figure 6 shows the $\|\cdot\|_{1,\Omega}$-norm of the solution as a function of the parameter λ. We see that there is a **principal solution branch** which has a left-winding **fold point** for some critical parameter value λ^* (actually $\lambda* \approx 6.8$ for the Bratu problem). The principal solution branch up to that critical value represents the physically **stable branch** whereas the upper part of the principal solution branch corresponds to

the physically **unstable branch**. Moreover, on the unstable branch, **primary bifurcation** from the principal branch as well as **secondary bifurcation** may occur.

6.1 Newton continuation methods

6.1.1 Introduction

If we discretize the Bratu problem (1.315) by finite differences or finite elements, it takes the form of a **parameter dependent system of nonlinear equations** in Euclidean space $\mathbb{R}^n$:
Given a continuously differentiable function $F : D \times I \to \mathbb{R}^n$, $D \subset \mathbb{R}^n$, $I \subset \mathbb{R}$, find $(x, \lambda) \in D \times I$ such that

$$F(x, \lambda) \;=\; 0 \;. \tag{1.316}$$

Assume that (x^*, λ^*) is the unique solution of (1.316) in $D \times I$ and that there exists a neighborhood $U(x^*) \times I(\lambda^*) \subset D \times I$ such that $F_x(x, \lambda) \in \mathbb{R}^{n \times n}$ is non-singular for all $(x, \lambda) \in U(x^*) \times I(\lambda^*)$. Then, the **implicit function theorem** asserts the existence of a continuously differentiable function

$$\overline{x} : I(\lambda^*) \to U(x^*) \;,$$

called the **homotopy path**, such that $\overline{x}(\lambda^*) = x^*$ and

$$F(\overline{x}, \lambda) \;=\; 0 \quad , \quad \lambda \in I(x^*) \;. \tag{1.317}$$

Differentiation with respect to λ results in a linearly implicit ODE, called the **Davidenko differential equation**

$$F_x \overline{x}_\lambda \;+\; F_\lambda \;=\; 0 \quad , \quad \lambda \in I(x^*) \;. \tag{1.318}$$

By introducing the **augmented variable**

$$y \;:=\; (x, \lambda) \in \mathbb{R}^{n+1} \;, \tag{1.319}$$

equation (1.316) can be rewritten as

$$F(y) \;=\; 0 \;, \tag{1.320}$$

whereas the Davidenko equation takes the form

$$F'(y)t(y) \;=\; 0 \tag{1.321}$$

with the path $t(y)$ being uniquely defined in a neighborhood of $y^* = (x^*, \lambda^*)$ up to some appropriate normalization (e.g., $\|t\| = 1$), provided

$$\text{rank } F'(y^*) \;=\; n \quad \Longleftrightarrow \quad \dim \;\ker F'(y^*) \;=\; 1 \;. \tag{1.322}$$

However, if for some $k \geq 1$ there holds

$$\text{rank } F'(y^*) \;=\; n-k \quad \Longleftrightarrow \quad \dim \ \ker F'(y^*) \;=\; k+1 \;, \tag{1.323}$$

the point $y^* = (x^*, \lambda^*)$ is a **critical point**.
In particular, in case $k = 1$, the point y^* is said to be a **simple bifurcation point**.
A special role is played by **turning points** or **fold points** which occur for $k = 0$, if

$$\text{rank } F'(y^*) \;=\; n \quad \text{and} \quad \text{rank } F_x(x^*, \lambda^*) \;=\; n-1 \;. \tag{1.324}$$

As far as the **numerical solution** of (1.316) is concerned, from a theoretical point of view one might be tempted to solve the Davidenko equation (1.318) by an appropriate numerical integrator for implicit ODEs. This, however, is not suitable, since

- it usually requires second order information in terms of F_{xx} , $F_{\lambda x}$ whose computation can be computationally costly,

and more importantly,

- it does not enforce directly the solution of (1.316) so that due to error propagation the computed solution path may drift away from the true solution path.

Therefore, in the sequel we will focus on **discrete continuation methods**.

6.1.2 Classification of continuation methods

For the solution of the parameter dependent problem (1.316) we subdivide the interval $I := [0, L]$ into subintervals by means of the partition

$$0 =: \lambda_0 < \lambda_1 < \ldots < \lambda_{N-1} < \lambda_N := L \;,$$

and consider the local problems

$$F(x, \lambda_\nu) \;=\; 0 \quad , \quad 0 \leq \nu \leq N \;. \tag{1.325}$$

The solution of (1.325) requires a good initial guess which will be provided by some appropriately chosen **prediction method**. A related important issue is an **adaptive selection of the steplengths**

$$\Delta\lambda_\nu \;:=\; \lambda_{\nu+1} - \lambda_\nu \quad , \quad 0 \leq \nu \leq N-1 \;.$$

Both issues will be addressed within an **affine covariant theory**.

denoting the solution of (1.325) by $\overline{x}(\lambda), \lambda_\nu \leq \lambda \leq \lambda_{\nu+1}$, it is, of course, natural to start from $\overline{x}(\lambda_\nu)$. However, there are different ways to construct a **prediction path** $\hat{x}(\lambda), \lambda_\nu \leq \lambda \leq \lambda_{\nu+1}$, emanating from $\overline{x}(\lambda_\nu)$. In general, a **continuation method** defined via a **prediction path** $\hat{x}(\lambda)$ is said to be of order p, if there exists a positive constant η_p such that

$$\|\overline{x}(\lambda) - \hat{x}(\lambda)\| \;\leq\; \eta_p \, \Delta\lambda^p \;, \tag{1.326}$$

where $\Delta\lambda := \lambda - \lambda_\nu$.
In the sequel, as the most important examples for prediction pathes we will consider

- the classical continuation method,
- the tangent continuation method,
- the standard and the partial standard embedding,
- the polynomial continuation method.

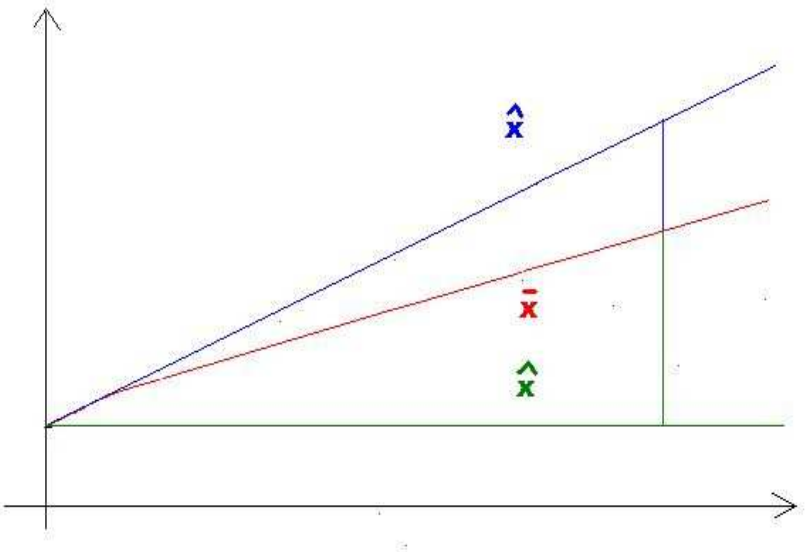

Figure 7: Classical (green) and tangent (blue) continuation

(i) Classical continuation

The most simple prediction path is the constant path

$$\hat{x}(\lambda) \;:=\; \overline{x}(\lambda_\nu) \quad , \quad \lambda_\nu \le \lambda \le \lambda_{\nu+1} \; . \tag{1.327}$$

Obviously, we have

$$\|\overline{x}(\lambda) - \hat{x}(\lambda)\| \;\le\; \|\overline{x}(\lambda) - \overline{x}(\lambda_\nu)\| \;\le\; \Delta\lambda \max_{s\in[\lambda_\nu,\lambda_{\nu+1}]} \|\overline{x}'(s)\| \; .$$

Hence, the method is of first order with the order coefficient

$$\eta_1 \;:=\; \max_{s\in[\lambda_\nu,\lambda_{\nu+1}]} \|\overline{x}'(s)\| \; .$$

(ii) Tangent continuation

An alternative way to obtain a prediction path is to apply the explicit Euler method to the Davidenko equation (1.318):

$$\hat{x}(\lambda) \;:=\; \overline{x}(\lambda_\nu) \;+\; (\lambda - \lambda_\nu)\, \overline{x}'(\lambda_\nu) \quad , \quad \lambda_\nu \le \lambda \le \lambda_{\nu+1} \; . \tag{1.328}$$

Therefore, the **tangent continuation** is also referred to as the **Euler continuation** or **method of incremental load**. Figure 7 shows both the classical and the tangent continuation method.
As far as the order is concerned, we have

$$\|\overline{x}(\lambda) - \hat{x}(\lambda)\| \;\le\; \|\overline{x}(\lambda) - \overline{x}(\lambda_\nu) - \lambda\, \overline{x}'(\lambda_\nu)\| \;\le\; \frac{1}{2}\, \Delta\lambda^2 \max_{s\in[\lambda_\nu,\lambda_{\nu+1}]} \|\overline{x}''(s)\| \; .$$

Hence, the method is of second order with the order coefficient

$$\eta_2 \;:=\; \frac{1}{2} \max_{s\in[\lambda_\nu,\lambda_{\nu+1}]} \|\overline{x}''(s)\| \; .$$

(iii) Standard embedding and partial standard embedding

Another way to obtain a prediction path is to consider the **embedding**

$$F_\nu(x,\lambda) \;:=\; F(x) \;-\; (1-\lambda)\, F(x_\nu) \quad , \quad \lambda_\nu \le \lambda \le \lambda_{\nu+1} \; . \tag{1.329}$$

However, this method does not exploit any structure of F. Therefore, a better way is ro select only a component of the mapping which leads to the so-called **partial standard embedding**

$$\overline{F}_\nu(x,\lambda) \;:=\; PF(x) - P^\perp\Big(F(x) - (1-\lambda)\, F(x_\nu)\Big) \; , \; \lambda_\nu \le \lambda \le \lambda_{\nu+1} \; , \tag{1.330}$$

where P is an appropriate orthogonal projection.

Lemma 6.1 Classical and tangent continuation based on partial standard embedding

Assume that $F : D \times I \to \mathbb{R}^n$ is continuously differentiable in the first argument with regular Jacobian F' and that the following **affine covariant Lipschitz condition** holds true

$$\|F'(\hat{x}(\lambda))^{-1}\Big(F'(x) - F'(\hat{x}(\lambda)\Big)\| \leq \hat{\omega}\ \|x - \hat{x}(\lambda)\|\ ,\ x, \hat{x}(\lambda) \in D\ ,\ \lambda \in I \tag{1.331}$$

Then, the order coefficient for the classical continuation method is

$$\eta_1\ =\ \max_{\lambda \in I} \|F'(\overline{x}(\lambda))^{-1} P^{\perp} F(x_\nu)\|\ , \tag{1.332}$$

whereas that one for the tangent continuation method is close to

$$\eta_2\ :=\ \frac{1}{2}\ \hat{\omega}\ \eta_1^2\ . \tag{1.333}$$

Proof. The partial derivatives of the map $\overline{F}_\nu$ can be easily computed

$$\frac{\partial}{\partial x}\ \overline{F}_\nu(x, \lambda) = \frac{\partial}{\partial x}\ F(x) = F'(x) \quad , \quad \frac{\partial}{\partial \lambda}\ \overline{F}_\nu(x, \lambda) = P^{\perp} F(x_\nu)\ ,$$

and hence,

$$\overline{x}'(\lambda)\ =\ -\ F'(\overline{x}(\lambda))^{-1} P^{\perp} F(x_\nu)\ .$$

This readily gives (1.332). For the tangent continuation, we must invoke the Lipschitz condition

$$\begin{aligned}
&\|\overline{x}'(\lambda) - \overline{x}'(\lambda_\nu)\|\ =\ \|\Big(F'(\overline{x}(\lambda))^{-1} - F'(\overline{x}(\lambda_\nu))^{-1}\Big) P^{\perp} F(x_\nu)\|\ \leq \\
&\leq\ \|F'(\hat{x}(\lambda_\nu))^{-1}\Big(F'(\overline{x}(\lambda))\ -\ F'(\overline{x}(\lambda_\nu))\Big)\|\ \|\overline{x}'(\lambda)\|\ \leq \\
&\leq\ \hat{\omega}\ \|\overline{x}(\lambda) - \overline{x}(\lambda_\nu)\|\ \eta_1\ \leq\ \hat{\omega}\ \eta_1^2\ \lambda\ .
\end{aligned}$$

(iv) Polynomial continuation

We distinguish between extrapolation by Lagrange and by Hermite interpolation.

$(iv)_1$ Lagrange extrapolation

We assume that for some $q > 0$ the data

$$\overline{x}(\lambda_\ell) \quad , \quad \nu - q \leq \ell \leq \nu$$

is available. Then, in terms of the fundamental Lagrange polynomials $L_q^\ell(\cdot)$, the prediction path is given by the interpolating polynomial

$$\hat{x}_q(\lambda) \;:=\; \sum_{\ell=\nu-q}^{\nu} \overline{x}(\lambda_\ell \; L_q^\ell(\lambda) \; . \tag{1.334}$$

Standard error estimates give

$$\|\overline{x}(\lambda) - \hat{x}_q(\lambda)\| \;\leq\; C_{q+1} \; \varphi(\lambda) \; , \tag{1.335}$$

where

$$\varphi(\lambda) \;:=\; \prod_{\ell=\nu-q}^{\nu} (\lambda - \lambda_\ell) \; .$$

(iv)$_2$ Hermite extrapolation

Here, we assume that we are given the data

$$\overline{x}(\lambda_\ell) \; , \; \overline{x}'(\lambda_\ell) \quad , \quad \nu - q \leq \ell \leq \nu \; .$$

We define the prediction path $\hat{x}_q(\lambda)$ as the associated Hermite polynomial and obtain

$$\|\overline{x}(\lambda) - \hat{x}_q(\lambda)\| \;\leq\; \overline{C}_{q+1} \; \overline{\varphi}(\lambda) \; , \tag{1.336}$$

where

$$\overline{\varphi}(\lambda) \;:=\; \prod_{\ell=\nu-q}^{\nu} (\lambda - \lambda_\ell)^2 \; .$$

6.1.3 Affine covariant correction method

Once we have computed a prediction path $\hat{x}(\lambda), \lambda_\nu \leq \lambda \leq \lambda_{\nu+1}$, we choose the predicted value $x^0 := \hat{x}(\lambda_{\nu+1})$ as an initial guess for a **correction method** to compute an approximation of $x^* := \overline{x}(\lambda_{\nu+1})$. We will study the **ordinary Newton method** with a new Jacobian at each iterate. Applying the affine covariant version of the Newton-Kantorovich theorem, we get the following result.

Theorem 6.1 Convergence of the corrector

Assume that $F : D \times I \to \mathbb{R}^n$ is continuously differentiable with nonsingular Jacobian $F_x(x, \lambda), (x, \lambda) \in D \times I$. Further, suppose that there exists a unique homotopy path $\overline{x}(\lambda)$ and that the **affine covariant Lipschitz condition**

$$\|F_x(\hat{x}(\lambda), \lambda)^{-1}\Big(F_x(y, \lambda) - F_x(x, \lambda)\Big)\| \leq \hat{\omega}_0 \; \|y - x\| \; , \; x, y \in D \; , \; \lambda \in I \tag{1.337}$$

is satisfied, where $\hat{x}(\lambda)$ is a prediction method of order p (cf. (1.326)). Then, for all step sizes

$$\Delta\lambda_\nu \;\leq\; \Delta\lambda_{max} \;:=\; \Big(\frac{\sqrt{2}-1}{\hat{\omega}_0\;\eta_p}\Big)^{1/p}\;, \tag{1.338}$$

the ordinary Newton method with initial guess $\hat{x}(\lambda_{\nu+1})$ converges to the solution point $\overline{x}(\lambda_{\nu+1})$.

Proof. For the ease of exposition, we write λ instead of $\Delta\lambda$. The affine covariant Newton-Kantorovich theorem requires

$$\|\Delta x^0(\lambda)\|\hat{\omega}_0 \;\leq\; \frac{1}{2}\;. \tag{1.339}$$

Applying the Lipschitz condition (1.337), by straightforward computation we find

$$\|\Delta x^0(\lambda)\| = \|F_x(\hat{x}(\lambda),\lambda)^{-1}F(\hat{x}(\lambda),\lambda)\| = \|F_x(\hat{x},\lambda)^{-1}\Big(F(\hat{x},\lambda)-F(\overline{x},\lambda)\Big)\| \;=$$

$$\|F_x(\hat{x},\lambda)^{-1}\int_0^1 F_x(\overline{x}+t(\hat{x}-\overline{x}),\lambda)(\hat{x}-\overline{x})\;dt\| \leq \|\hat{x}-\overline{x})\|\Big(1+\frac{1}{2}\;\hat{\omega}_0\;\|\hat{x}-\overline{x}\|\Big)\;.$$

Observing (1.326), we deduce

$$\|\Delta x^0(\lambda)\| \leq \eta_p\lambda^p\Big(1+\frac{1}{2}\;\hat{\omega}_0\;\eta_p\lambda^p\Big) \;=:\; \alpha(\lambda)\;. \tag{1.340}$$

Consequently, this leads to the requirement

$$\hat{\omega}_0\;\eta_p\lambda^p\Big(1+\frac{1}{2}\;\hat{\omega}_0\;\eta_p\lambda^p\Big) \;\leq\; \frac{1}{2}\;,$$

which is equivalent to

$$\hat{\omega}_0\;\eta_p\lambda^p \;\leq\; \sqrt{2}-1\;.$$

•

6.1.4 Adaptive stepsize control

For the practical application of the theoretical convergence results we have to replace the theoretical quantities $\hat{\omega}_0$ and η_p by computationally available lower bounds $[\hat{\omega}_0]$ and $[\eta_p]$ thus resulting in the **stepsize estimate**

$$[\Delta\lambda_{max}] \;:=\; \Big(\frac{\sqrt{2}-1}{[\hat{\omega}_0]\;[\eta_p]}\Big)^{1/p} \;\geq\; \Delta\lambda_{max}\;. \tag{1.341}$$

Since there might be a substantial overestimation, we need again a **prediction strategy** and a **correction strategy**.

As far as the **correction strategy** is concerned, let us assume that for $\lambda_{\nu+1}$we already know the first contraction factor

$$\Theta_0(\lambda) \;:=\; \frac{\|\Delta x^1(\lambda)\|}{\|\Delta x^0(\lambda)\|} \;.$$

The convergence analysis of the affine covariant Newton method yields

$$\Theta_0(\lambda) \;\leq\; \frac{1}{2}\;\hat{\omega}_0\;\|\Delta x^0(\lambda)\| \;. \tag{1.342}$$

Hence, inserting (1.340) gives us

$$\Theta_0(\lambda) \;\leq\; \frac{1}{2}\hat{\omega}_0\;\eta_p\Delta\lambda^p \;,$$

which leads to

$$\hat{\omega}_0\;\eta_p\Delta\lambda^p \;\geq\; g(\Theta_0(\lambda)) \;,$$

where

$$g(\Theta) \;:=\; \sqrt{1+4\Theta} \;-\; 1 \;.$$

From this, we get the **a posteriori estimate**

$$[\hat{\omega}_0\;\eta_p] \;:=\; \frac{g(\theta_0(\lambda))}{\Delta\lambda^p} \;\leq\; \hat{\omega}_0\;\eta_p \;,$$

and the associated **stepsize estimate**

$$[\Delta\lambda_{max}] \;:=\; \Big(\frac{g(\overline{\Theta})}{[\hat{\omega}_0\;\eta_p]}\Big)^{1/p} \quad , \quad \overline{\Theta} = \frac{1}{4} \;.$$

Denoting by $\Delta\lambda_\nu$ the stepsize associated with the computed value of Θ_0 and by $\Delta\lambda'_\nu$ corresponding to $\overline{\Theta} = \frac{1}{4}$, we arrive at the **stepsize correction**

$$\Delta\lambda'_\nu \;:=\; \Big(\frac{g(\overline{\Theta})}{g(\Theta_0)}\Big)^{1/p}\;\Delta\lambda_\nu \;. \tag{1.343}$$

Remark: If the termination criterion detects some Θ_k such that $\Theta_k > \frac{1}{2}$, the last continuation step has to be repeated with

$$\Delta\lambda'_\nu \;:=\; \Big(\frac{g(\overline{\Theta})}{g(\Theta_k)}\Big)^{1/p}\;\Delta\lambda_\nu \;, \tag{1.344}$$

which gives rise to a reduction, since

$$[\Delta\lambda_{max}] \;<\; \Big(\frac{\sqrt{2}-1}{\sqrt{3}-1}\Big)^{1/p}\;\Delta\lambda_\nu \;\approx\; 0.57^p\;\Delta\lambda_\nu \;.$$

Whereas a posteriori estimates lead to correction strategies, **a priori estimates** allow us to derive **prediction methods**.
We note that (1.342) gives us the lower bound

$$[\hat{\omega}_0] \;:=\; \frac{2\Theta_0(\lambda_\nu)}{\|\Delta x^0(\lambda_\nu)} \;\leq\; \hat{\omega}_0 \;,$$

whereas the definition (1.330) of the order of a prediction method implies

$$[\eta_p] \;:=\; \frac{\|\hat{x}(\lambda_\nu) - \overline{x}(\lambda_\nu)\|}{|\Delta\lambda_{\nu-1}|^p} \;\leq\; \eta_p \;.$$

Using the preceding quantities in (1.341), we arrive at the **stepsize prediction**

$$\Delta\lambda_\nu^0 \;:=\; \Big(\frac{\|\Delta x^0(\lambda_\nu)\|}{\|\hat{x}(\lambda_\nu) - \overline{x}(\lambda_\nu)\|}\frac{g(\overline{\Theta})}{2\Theta_0}\Big)^{1/p} \Delta\lambda_{\nu-1} \;. \tag{1.345}$$

Remark: The prediction strategy (1.345) is **robust** with respect to the accuracy of $\overline{x}$: Even if only a single Newton step is performed, i.e.,

$$\overline{x} \;=\; \hat{x}(\lambda_\nu) \;+\; \Delta x^0(\lambda_\nu) \;,$$

the prediction takes the form

$$\Delta\lambda_\nu^0 \;:=\; \Big(\frac{g(\overline{\Theta})}{2\Theta_0}\Big)^{1/p} \Delta\lambda_{\nu-1} \;,$$

which in lights of (1.343) still is a reasonable estimate.
Only in the **nearly linear case**

$$\Theta_0 \;\leq\; \Theta_{min} \;\ll\; 1 \;,$$

the estimate (1.345) should be replaced by

$$\Delta\lambda_\nu^0 \;:=\; \Big(\frac{g(\overline{\Theta})}{2\Theta_{min}}\Big)^{1/p} \Delta\lambda_{\nu-1} \;. \tag{1.346}$$

6.2 Augmented systems for critical points

We assume that y^* is a perfect or unperturbed singularity of order $k \geq 1$ such that

$$F(y^*) \;=\; 0 \quad , \quad \text{rank } F'(y^*) \;=\; n-k \;. \tag{1.347}$$

For notational convenience, we set $A := F'(y^*)$ and refer to

$$\mathcal{N}(A) \;:=\; \ker A \quad , \quad \mathcal{R}^{\perp}(A) \tag{1.348}$$

with

$$\dim \mathcal{N}(A) \;=\; k+1 \quad , \quad \dim \mathcal{R}^{\perp}(A) \;=\; k$$

as the nullspace and corange of A. We further introduce the projectors

$$P \;:=\; A^+A \quad , \quad \overline{P} \;:=\; AA^+ \tag{1.349}$$

and recall that $P^{\perp}$ projects onto $\mathcal{N}(A)$, whereas $\overline{P}^{\perp}$ projects onto $\mathcal{R}^{\perp}(A)$.
We consider the natural splitting

$$y \;=\; y^* + v + w \quad , \quad w \;:=\; P(y-y^*) \quad , \quad v \;:=\; P^{\perp}(y-y^*) \;. \tag{1.350}$$

Then, in view of (1.347) the implicit function theorem asserts the existence of a function $w^* = w^*(v)$ such that

$$\overline{P}F(y^*+v+w) \;=\; 0 \quad \Longleftrightarrow \quad w \;=\; w^*(v) \;. \tag{1.351}$$

Replacing w by w^* gives rise to the reduced system

$$f(v) \;:\; \overline{P}^{\perp}F(y^*+v+w^*(v)) \;=\; 0 \;, \tag{1.352}$$

which is known as the Lyapunov-Schmidt reduction.
For computational purposes, we choose orthogonal bases of the nullspace and the corange according to

$$\mathcal{N}(A) \;:=\; \text{span}(t_1,...,t_{k+1}) \quad , \quad \mathcal{R}^{\perp}(A) \;:=\; \text{span}(z_1,...,z_k) \;.$$

In terms of the matrices

$$t \;:=\; [t_1,...,t_{k+1}] \quad , \quad z \;:=\; [z_1,...,z_k] \;,$$

we obviously have

$$\begin{aligned} At &= 0\ ,\ t^T t = I_{k+1}\ ,\ P^{\perp} = tt^T\ , \\ A^T z &= 0\ ,\ z^T z = I_k\ ,\ \overline{P}^{\perp} = zz^T\ . \end{aligned} \tag{1.353}$$

If we now introduce local coordinates by means of

$$v = t\xi = \sum_{i=1}^{k+1} \xi_i t_i \quad , \quad f(v) = z\gamma = \sum_{j=1}^{k} \gamma_j z_j\ ,$$

in terms of the function $\gamma : \mathbb{R}^{k+1} \to \mathbb{R}^k$, the reduced system (1.351) can be written as

$$\gamma(\xi) := z^T f(t\xi) = z^T F(y^* + t\xi + w^*(t\xi)) = 0\ . \tag{1.354}$$

For the actual computation of singularities, we need higher order derivatives which are provided by the following result.

Lemma 6.2 Higher order derivatives w.r.t. the Lyapunov-Schmidt reduction

Assume for simplicity that $y^* = 0$. Then, with the notations introduced above, for $a_i \in \mathbb{R}^{k+1}, 1 \le i \le 3$, there holds

$$\frac{d\gamma}{d\xi}(0)a_1 = 0\ , \tag{1.355}$$

$$\frac{d^2\gamma}{d\xi^2}(0)[a_1, a_2] = z^T F''[ta_1, ta_2]\ , \tag{1.356}$$

$$\begin{aligned} \frac{d^3\gamma}{d\xi^3}(0)[a_1, a_2, a_3] &= z^T F'''[ta_1, ta_2, ta_3] - \\ &- z^T F''[ta_1, A^+ F''[ta_2, ta_3]] - \\ &- z^T F''[ta_2, A^+ F''[ta_3, ta_1]] - \\ &- z^T F''[ta_3, A^+ F''[ta_1, ta_2]]\ . \end{aligned} \tag{1.357}$$

In view of the preceding result, it is sufficient to consider the contact equivalence class

$$\Gamma(g) := \{\ \gamma(\xi) = \beta(\xi) g(h(\xi))\ \}\ , \tag{1.358}$$

where $g \in P^{k+1+q}(\xi)$ with q denoting the codimension of $\mathcal{N}(A)$ and β, h are C^∞-diffeomorphisms with $h(0) = 0$. The functions g are called polynomial germs.

For instance, in case of a simple bifurcation, i.e., $k = 1, q = 1$, we have

$$g(\xi) = \xi_1^2 - \xi_2^2\ , \tag{1.359}$$

whereas for an asymmetric cusp, i.e., $k = 1, q = 2$,we obtain

$$g(\xi) \;=\; \xi_1^2 \;-\; \xi_2^3 \;. \tag{1.360}$$

In order to allow for imperfect or unfolded singularities, we have to consider the perturbed germs

$$G(\xi, \alpha) \;:=\; g(\xi) \;+\; p(\xi, \alpha) \;, \tag{1.361}$$

where $p(\xi, \alpha) \in P^{q-1}(\xi)$ is a polynomial perturbation with $\alpha \in \mathbb{R}^q$ denoting the unfolding parameters.
For instance, for a simple bifurcation

$$G(\xi, \alpha) \;=\; \xi_1^2 \;-\; \xi_2^2 \;+\; \alpha \;, \tag{1.362}$$

and for an asymmetric cusp

$$G(\xi, \alpha) \;=\; \xi_1^2 \;-\; \xi_2^3 \;+\; \alpha_1 \;+\; \alpha_2 \xi_2 \;. \tag{1.363}$$

In particular, if we have

$$G(h(0), \alpha) \;=\; G(o, \alpha) \;=\; p(0, \alpha) \;\neq\; 0 \;,$$

the reduced system (1.354) has to be replaced by

$$z^T F(y^*) \;=\; p(0, \alpha) \;.$$

These k equations together with the $n - k$ equations $\overline{P}F = 0$ then give rise to the n equations

$$F(y^*) \;=\; zp(0, \alpha) \tag{1.364}$$

in the $(k + 1)n + q + 1$ unknowns (y, z, α).
Normalizing the k basis functions $z_j, 1 \leq j \leq k$, in case of a simple bifurcation we arrive at the augmented system

$$F'(y)^T z \;=\; 0 \;, \tag{1.365}$$

$$F(y) + \alpha z \;=\; 0 \;, \tag{1.366}$$

$$\frac{1}{2}\,(z^T z - 1) \;=\; 0 \;, \tag{1.367}$$

which represents $2n + 2$ equations in the $2n + 2$ unknowns (y, z, α).
Including second order information, we can show the existence of two local branch directions.

Theorem 6.2 Existence of two local branch directions

Let y^* be a simple bifurcation and assume that $F \in C^k, k \geq 3$, and that

$$z^T F''(y^*)[t,t] \tag{1.368}$$

is nondegenerate. Then, in a neighborhood of y^*, the solution set of $F + \alpha z = 0$ consists of two one-dimensional C^{k-2}-branches γ_1, γ_2 such that

$$\begin{aligned} \gamma_i(0) &= y^* \ , \ 1 \leq i \leq 2 \ , \\ \mathcal{N} &= \{\dot{\gamma}_1(0), \dot{\gamma}_2(0)\} \ , \\ z^T F''(y^*)\Big[\dot{\gamma}_1(0), \dot{\gamma}_2(0)\Big] &= 0 \ . \end{aligned}$$

6.3 Newton method for simple bifurcations

For the augmented system (1.365), the extended Jacobian has the following block structure

$$J(y,z,\alpha) = \begin{pmatrix} C & A^T & 0 \\ A & \alpha I_n & z \\ 0 & z^T & 0 \end{pmatrix} \ ,$$

where

$$C := (F'(y)^T z)' = \sum_{i=1}^{n} f_i''(y) z_i \quad , \quad A := F'(y) \ .$$

Theorem 6.3 Properties of the extended Jacobian

At a simple bifurcation point y^* with sufficiently small perturbation parameter α^*, the extended Jacobian $J(y^*, z^*, \alpha^*)$ is nonsingular.

As a consequence of Theorem 6.3, the ordinary Newton method

$$\begin{pmatrix} C & A^T & 0 \\ A & \alpha I_n & z \\ 0 & z^T & 0 \end{pmatrix} \begin{pmatrix} \Delta y \\ \Delta z \\ \Delta \alpha \end{pmatrix} = - \begin{pmatrix} (F')^T z \\ F + \alpha z \\ \frac{1}{2}(z^T z - 1) \end{pmatrix} \tag{1.369}$$

is well-defined in a neighborhood of y^*.
Instead of (1.369), replacing $J(y,z,\alpha)$ by $J(y,z,o)$ and A by $\tilde{A} \approx F'(y^*)$, we consider the Newton-like method

$$\begin{pmatrix} C & \tilde{A}^T & 0 \\ \tilde{A} & 0 & z \\ 0 & z^T & 0 \end{pmatrix} \begin{pmatrix} \Delta y \\ \Delta z \\ \Delta \alpha \end{pmatrix} = - \begin{pmatrix} (F')^T z \\ F + \alpha z \\ \frac{1}{2}(z^T z - 1) \end{pmatrix} \ , \tag{1.370}$$

which is easier to solve.
In particular, a structure preserving algorithm for the solution of (1.370) makes use of the following QR decomposition of $A = F'(y)$

$$A \;=\; Q \begin{pmatrix} R & S \\ 0 & \varepsilon^T \end{pmatrix} \Pi^T \;,$$

where Q is an orthogonal $n \times n$ matrix, R is an upper triangular $(n-1)\times(n-1)$ matrix, S is an $(n-1) \times 2$ matrix, $\varepsilon \in \mathbb{R}^2$, and Π is an $(n+1) \times (n+1)$ permutation matrix.
For y close to y^*, the matrix R is nonsingular and the vector ε is small. Hence, we may choose

$$\tilde{A} \;=\; Q \begin{pmatrix} R & S \\ 0 & 0 \end{pmatrix} \Pi^T \;. \tag{1.371}$$

Using (1.371) in (1.370), suggests the partitioning

$$\begin{aligned}
\hat{C} \;:=\; \Pi^T C \Pi \;&=\; \begin{pmatrix} C_{11} & C_{12} \\ C_{12}^T & C_{22} \end{pmatrix} \quad , \quad C_{22} \in \mathbb{R}^{2\times 2} \;, \\
\overline{z} \;:=\; Q^T z \;&=\; \begin{pmatrix} w \\ \zeta \end{pmatrix} \quad , \quad w \in \mathbb{R}^{n-1} \;,\; \zeta \in \mathbb{R} \;, \\
\Pi^T \Delta y \;&=\; \begin{pmatrix} \Delta u \\ \Delta v \end{pmatrix} \quad , \quad \Delta u \in \mathbb{R}^{n-1} \;,\; \Delta v \in \mathbb{R}^2 \;, \\
\Delta \overline{z} \;=\; Q^T \Delta z \;&=\; \begin{pmatrix} \Delta w \\ \Delta \zeta \end{pmatrix} \quad , \quad \Delta w \in \mathbb{R}^{n+1} \;,\; \Delta \zeta \in \mathbb{R} \;, \\
\Pi^T A^T z \;&=\; \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \quad , \quad f_1 \in \mathbb{R}^{n-1} \;,\; f_2 \in \mathbb{R}^2 \;, \\
Q^T (F + \alpha z) \;&=\; \begin{pmatrix} g_1 \\ g_2 \end{pmatrix} \quad , \quad g_1 \in \mathbb{R}^{n-1} \;,\; g_2 \in \mathbb{R} \;, \\
h \;&:=\; \frac{1}{2} \left(z^T z - 1 \right) \;,
\end{aligned}$$

which leads to the linear system

$$\begin{pmatrix} C_{11} & C_{12} & R^T & 0 & 0 \\ C_{12}^T & C_{22} & S^T & 0 & 0 \\ R & S & 0 & 0 & w \\ 0 & 0 & 0 & 0 & \zeta \\ 0 & 0 & w^T & \zeta & 0 \end{pmatrix} \begin{pmatrix} \Delta u \\ \Delta v \\ \Delta w \\ \Delta \zeta \\ \Delta \alpha \end{pmatrix} \;=\; - \begin{pmatrix} f_1 \\ f_2 \\ g_1 \\ g_2 \\ h \end{pmatrix} \;.$$