2 Systems of Equations: Local Newton Methods

This chapter deals with the numerical solution of systems of nonlinear equations with finite, possibly large dimension n. The term *local* Newton methods refers to the situation that—only throughout this chapter—'sufficiently good' initial guesses of the solution are assumed to be at hand. Special attention is paid to the issue of how to recognize—in a computationally cheap way—whether a given initial guess x^0 is 'sufficiently good'. As it turns out, different affine invariant Lipschitz conditions, which have been introduced in Section 1.2.2, lead to different characterizations of local convergence domains in terms of error oriented norms, residual norms, or energy norms and convex functionals, which, in turn, give rise to corresponding variants of Newton algorithms.

We give three different, strictly affine invariant convergence analyses for the cases of affine covariant (error oriented) Newton methods (Section 2.1), affine contravariant (residual based) Newton methods (Section 2.2), and affine conjugate Newton methods for convex optimization (Section 2.3). Details are worked out for ordinary Newton algorithms, simplified Newton algorithms, and inexact Newton algorithms—synoptically for the three affine invariance classes. Moreover, affine covariance appears as associated with Broyden's 'good' quasi-Newton method, whereas affine contravariance corresponds to Broyden's 'bad' quasi-Newton method.

2.1 Error Oriented Algorithms

A convergence analysis for any error oriented algorithm of Newton type will start from *affine covariant* Lipschitz conditions of the kind (1.7) and lead to results in the space of the iterates only. The behavior of the residuals will be ignored. For actual computation, scaling of any arising norms of Newton corrections is tacitly assumed.

2.1.1 Ordinary Newton method

Consider the ordinary Newton method in the notation

P. Deuflhard, *Newton Methods for Nonlinear Problems: Affine Invariance and Adaptive Algorithms*, Springer Series in Computational Mathematics 35,
DOI 10.1007/978-3-642-23899-4_2, © Springer-Verlag Berlin Heidelberg 2011

$$F'(x^k)\Delta x^k = -F(x^k)\,,\ x^{k+1} = x^k + \Delta x^k\,, \quad k = 0, 1, \ldots\,. \tag{2.1}$$

Convergence analysis. Because of its fundamental importance, we begin with an affine covariant version of the classical 'Newton-Kantorovich theorem'. Only at this early stage we state the theorem in Banach spaces—well aware of the fact that a Banach space formulation is not directly applicable to numerical methods: in the numerical solution of nonlinear operator equations both function and derivative *approximations* must be taken into account. As a consequence, *inexact* Newton methods in Banach spaces are the correct theoretical frame to study convergence of algorithms—to be treated below in Sections 7.4, 8.1, and 8.3.

Theorem 2.1 *Let $F : D \to Y$ be a continuously Fréchet differentiable mapping with $D \subseteq X$ open and convex. For a starting point $x^0 \in D$ let $F'(x^0)$ be invertible. Assume that*

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

$$\|\, F'(x^0)^{-1}\,(F'(y) - F'(x))\, \| \leq \overline{\omega}_0 \|y - x\| \qquad x, y \in D\,, \tag{2.2}$$

$$h_0 := \alpha\overline{\omega}_0 \leq \tfrac{1}{2}\,, \tag{2.3}$$

$$\overline{S}(x^0, \rho_-) \subset D, \qquad \rho_- := \left(1 - \sqrt{1 - 2h_0}\right) / \,\overline{\omega}_0\,.$$

Then the sequence $\{x^k\}$ obtained from the ordinary Newton iteration is well-defined, remains in $\overline{S}(x^0, \rho_-)$, and converges to some x^* with $F(x^*) = 0$. For $h_0 < \frac{1}{2}$, the convergence is quadratic.

Proof. Rather than giving the classical 1948 proof [126] of L.V. Kantorovich, we here sketch an alternative affine covariant proof, which dates back to T. Yamamoto [202] in 1985.

The proof is by induction starting with $k = 0$. At iterate x^k, let the Fréchet derivative $F'(x^k)$ be invertible. Hence we may require the affine covariant Lipschitz condition

$$\|\, F'(x^k)^{-1}\big(F'(y) - F'(x)\big)\, \| \leq \overline{\omega}_k \|y - x\|$$

and define an associated first majorant

$$\overline{\omega}_k \|\Delta x^k\| \leq h_k\,.$$

As a preparation to show that with $F'(x^k)$ also the Fréchet derivative $F'(x^{k+1})$ is invertible, we define the operators

$$B_{k+1} := F'(x^k)^{-1}F'(x^{k+1})$$

and the associated second majorant

$$\|B_{k+1}^{-1}\| \leq \beta_{k+1} \,.$$

Consequently, for $k > 0$ we have the upper bound

$$\overline{\omega}_k \leq \beta_k \overline{\omega}_{k-1} \,.$$

By means of the operator perturbation lemma, we easily obtain

$$\beta_{k+1} = 1/(1 - h_k) \,. \tag{2.4}$$

Next, in order to exploit the above Lipschitz condition, we apply standard analytical techniques to obtain

$$\|x^{k+1} - x^k\| = \Big\|F'(x^k)^{-1} \int\limits_{s=0}^{1} \left[F'(x^{k-1} + s\Delta x^{k-1}) - F'(x^{k-1})\right] \Delta x^{k-1} ds\Big\| \,,$$

which implies

$$\overline{\omega}_k \|x^{k+1} - x^k\| \leq \tfrac{1}{2}\beta_k^2 h_{k-1}^2 =: h_k \,. \tag{2.5}$$

Combination of the two relations (2.5) and (2.4) then yields the single recursive equation

$$h_k = \frac{\frac{1}{2}h_{k-1}^2}{(1 - h_{k-1})^2} \,.$$

Herein contraction occurs, if

$$\frac{\frac{1}{2}h_0}{(1 - h_0)^2} < 1 \,,$$

which directly leads to $h_0 < \frac{1}{2}$. Under this assumption, the convergence is quadratic.

Things are more complicated for the limiting case $h_0 = \frac{1}{2}$, which requires extra consideration. In this case, we obtain $h_k = \frac{1}{2}$, $k = 1, 2, \ldots$, which implies

$$\beta_k = 2 \,, \qquad \overline{\omega}_k \leq 2^k \overline{\omega}_0 \,.$$

Insertion into the majorant inequality (2.5) then leads to

$$\lim_{k\to\infty} \overline{\omega}_k \|x^{k+1} - x^k\| \leq \lim_{k\to\infty} 2^k \overline{\omega}_0 \|x^{k+1} - x^k\| \leq \tfrac{1}{2} \,,$$

which verifies that

$$\lim_{k\to\infty} \|x^{k+1} - x^k\| \leq \lim_{k\to\infty} \frac{1}{2^{k-1}\overline{\omega}_0} = 0 \,.$$

In the latter case, the convergence is linear. □

Remark 2.1 If we define $t^{**} = 1 + \sqrt{1 - 2h_0}$, $\rho_+ = t^{**}/\overline{\omega_0}$, and assume that $\bar{S}(x^0, \rho_+) \subset D$, the solution x^* can be shown to be unique in $S(x^0, \rho_+)$. The corresponding proof is omitted here.

Bibliographical Note. The name 'Newton-Kantorovich theorem' has been coined, since historically L.V. Kantorovich was probably the first to prove convergence for Newton's method in Banach spaces. In 1939, he actually showed *linear* convergence (see [125]), but not earlier than 1948 he published his famous proof of *quadratic* convergence (see [126]). Even though this early theorem has already been phrased in affine covariant terms, nearly all (with few exceptions) of his later published versions lack this desirable property (see, e.g., the book by L.V. Kantorovich and G. Akhilov [127]). In 1949, I. Mysovskikh [155] presented an alternative meanwhile classical convergence theorem, which today is called 'Newton-Mysovskikh theorem'. That theorem was not affine invariant in any sense; the following Theorem 2.2 is an affine covariant version of it. In 1970, an interesting theorem for local convergence of Newton's method, already in affine covariant formulation, has been proved by H.B. Keller in [129], under the relaxed assumption of Hölder continuity of $F'(x)$—see Exercise 2.4. Since then a huge literature concerning different aspects of the classical theorems has unfolded, typically in not affine invariant form—compare, e.g., the monograph of F.A. Potra and V. Pták [171].

Not earlier than 1979, affine invariance as a subject of its own right within convergence analysis has been emphasized by P. Deuflhard and G. Heindl in [76]; this paper included an affine covariant (then called affine invariant) rephrasing of the classical Newton-Kantorovich and Newton-Mysovskikh theorem and permitted a new local convergence theorem for Gauss-Newton methods for nonlinear least squares problems—see Section 4.3.1. Also around that time H.G. Bock [29, 31, 32] adopted affine invariance and slightly weakened the Lipschitz condition in the affine covariant Newton-Mysovskikh theorem that had been given in [76]. Following the affine invariance message of [76], T. Yamamoto has introduced affine covariance into his subtle convergence estimates for Newton's method—see, e.g., his starting paper [202] and work thereafter. Later on, the earlier convergence theorem due to L.B. Rall [174], proved under the assumptions

$$\|F'(x^*)^{-1}\| \leq \beta_* \,, \ \|F'(x) - F'(y)\| \leq \gamma \|x - y\|$$

has been put into an affine covariant form by G. Bader [15]. For the improved variant of Rall's theorem due to W.C. Rheinboldt [176] see Exercise 2.5.

Throughout the subsequent convergence analysis for local Newton-type methods, we will mostly study extensions of the Newton-Mysovskikh theorem [155], which have turned out to be an extremely useful basis for the construction of algorithms. The subsequent Theorem 2.3 is the 'refined Newton-Mysovskikh theorem' due to P. Deuflhard and F.A. Potra [82], which has no classical predecessor, since it relies on affine covariance in its proof.

Next, we present an affine covariant Newton-Mysovskikh theorem. In what follows, we will return to the case of finite dimensional nonlinear equations, i.e. to $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$.

Theorem 2.2 *Let $F : D \to \mathbb{R}^n$ be a continuously differentiable mapping with $D \subset \mathbb{R}^n$ convex. Suppose that $F'(x)$ is invertible for each $x \in D$. Assume that the following affine covariant Lipschitz condition holds:*

$$\left\|F'(z)^{-1}\big(F'(y) - F'(x)\big)(y - x)\right\| \le \omega\|y - x\|^2$$

for collinear x, y, $z \in D$. For the initial guess x^0 assume that

$$h_0 := \omega\|\Delta x^0\| < 2 \tag{2.6}$$

and that $\bar{S}(x^0, \rho) \subset D$ for $\rho = \dfrac{\|\Delta x^0\|}{1 - \frac{1}{2}h_0}$.

Then the sequence $\{x^k\}$ of ordinary Newton iterates remains in $S(x^0, \rho)$ and converges to a solution $x^* \in \overline{S}(x^0, \rho)$. Moreover, the following error estimates hold

$$\|x^{k+1} - x^k\| \le \tfrac{1}{2}\omega\|x^k - x^{k-1}\|^2\,, \tag{2.7}$$

$$\|x^k - x^*\| \le \frac{\|x^k - x^{k+1}\|}{1 - \frac{1}{2}\omega\|x^k - x^{k+1}\|}\,. \tag{2.8}$$

Proof. First, the ordinary Newton iteration is used for k and $k - 1$:

$$\|\Delta x^k\| = \left\|F'(x^k)^{-1}\left[F(x^k) - \big(F(x^{k-1}) + F'(x^{k-1})\Delta x^{k-1}\big)\right]\right\|\,.$$

Application of the above Lipschitz condition yields

$$\|\Delta x^k\| \le \tfrac{1}{2}\omega\|\Delta x^{k-1}\|^2\,,$$

which is (2.7). For the purpose of repeated induction, introduce the following notation:

$$h_k := \omega\left\|\Delta x^k\right\|.$$

Multiplication of (2.7) by ω then leads to

$$h_k \le \tfrac{1}{2}h_{k-1}^2\,.$$

Contraction of the $\{h_k\}$ is obtained, if $h_0 < 2$ is assumed, which is just (2.6). From this, as in the proofs of the preceding theorems, we have $h_k < h_{k-1} < h_0 < 2$ so that there exists

$$\lim_{k\to\infty} h_k = 0\,.$$

A straightforward induction argument shows that

$$\|\Delta x^l\| \le \left(\tfrac{1}{2}h_k\right)^{l-k}\|\Delta x^k\| \text{ for } l \ge k\,.$$

Hence

$$\|x^{l+1} - x^k\| \le \|\Delta x^l\| + \dots + \|\Delta x^k\| \le \|\Delta x^k\| \sum_{j=0}^{\infty} \left(\tfrac{1}{2} h_k\right)^j = \frac{\|\Delta x^k\|}{1 - \frac{1}{2} h_k} .$$

The special case $k = 0$ implies that all Newton iterates remain in $S(x^0, \rho)$. Moreover, the results above show that $\{x^k\}$ is a Cauchy sequence, so it converges to some $x^* \in S(x^0, \rho)$. Taking the limit $l \to \infty$ on the previous estimate yields (2.8). Finally, with $\omega < \infty$ from (2.6) we have that x^* is a solution point. □

The following theorem has been named 'refined Newton-Mysovskikh theorem' in [82].

Theorem 2.3 *Let $F : D \to \mathbb{R}^n$ be a continuously differentiable mapping with $D \subset \mathbb{R}^n$ open and convex. Suppose that $F'(x)$ is invertible for each $x \in D$. Assume that the following affine covariant Lipschitz condition holds*

$$\left\|F'(x)^{-1}\big(F'(y) - F'(x)\big)(y - x)\right\| \le \omega \|y - x\|^2$$

for x, y, $\in D$. Let $F(x) = 0$ have a solution x^*.
For the initial guess x^0 assume that $\bar{S}(x^*, \|x^0 - x^*\|) \subset D$ and that

$$\omega \|x^0 - x^*\| < 2 . \tag{2.9}$$

Then the ordinary Newton iterates defined by (2.1) remain in the open ball $S(x^*, \|x^0 - x^*\|)$ and converge to x^* at an estimated rate

$$\|x^{k+1} - x^*\| \le \tfrac{1}{2} \omega \|x^k - x^*\|^2 . \tag{2.10}$$

Moreover, the solution x^* is unique in the open ball $S(x^*, 2/\omega)$.

Proof. We define $e_k := x^k - x^*$ and proceed for $\lambda \in [0, 1]$ as follows:

$$\begin{aligned}
\|x^k + \lambda \Delta x^k - x^*\| &= \|e_k - \lambda F'(x^k)^{-1}\left(F(x^k) - F(x^*)\right)\| \\
&= \|F'(x^k)^{-1}\big(\lambda\left(F(x^*) - F(x^k)\right) + F'(x^k) e_k\big)\| \\
&= \|(1 - \lambda) e_k + \lambda F'(x^k)^{-1} \int\limits_{s=0}^{1} (F'(x^k + s e_k) - F'(x^k)) e_k ds\| \\
&\le (1 - \lambda)\|e_k\| + \tfrac{\lambda}{2} \omega \|e_k\|^2 .
\end{aligned}$$

For the purpose of repeated induction assume that $\omega\|e_k\| \le \omega\|e_0\| < 2$ so that $x^k \in D$ is guaranteed. Then the above estimate can be continued to supply

$$\|x^k + \lambda \Delta x^k - x^*\| < (1 - \lambda)\|e_k\| + \lambda\|e_k\| = \|e_k\| \le \|e_0\| .$$

From this, any statement $x^k + \lambda \Delta x^k \notin S(x^*, \|x^0 - x^*\|)$ would lead to a contradiction. Hence, $x^{k+1} \in D$ and

$$\|e_{k+1}\| \le \tfrac{1}{2}\omega\|e_k\|^2\,,$$

which is just (2.10). In order to prove uniqueness in $S(x^*, 2/\omega)$, let $x^0 := x^{**}$ for some $x^{**} \neq x^*$ with $F(x^{**}) = 0$, which implies $x^1 = x^{**}$ as well. Insertion into (2.10) finally yields the contradiction

$$\|x^{**} - x^*\| \le \omega/2\ \|x^{**} - x^*\|^2 < \|x^{**} - x^*\|\,.$$

This completes the proof. □

In view of actual computation, we may combine the results of Theorem 2.2 and 2.3: if we require $h_k \le 1$, then contraction towards x^* shows up, since

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

Convergence monitor. We are now ready to exploit both convergence theorems for actual implementation of Newton's method. First, we define the contraction factors

$$\Theta_k := \frac{\|\Delta x^{k+1}\|}{\|\Delta x^k\|}\,,$$

which in terms of the unknown theoretical quantities h_k are known to satisfy

$$\Theta_k = \frac{h_{k+1}}{h_k} \le \tfrac{1}{2}h_k < 1\,. \tag{2.11}$$

Whenever $\Theta_k \ge 1$, then the ordinary Newton iteration is classified as 'not convergent'.

Computational Kantorovich estimates. Obviously, the assumption $h_0 \le 1$ implies $\Theta_0 \le 1/2$. We define the computationally available *a-posteriori* estimates

$$[h_k]_1 = 2\Theta_k \le h_k\ , \quad k = 0, 1, \ldots$$

and, recalling $h_{k+1} = \Theta_k h_k$ and shifting the index $k+1 \to k$, also corresponding *a-priori* estimates

$$[h_k] := \Theta_{k-1}[h_{k-1}]_1 = 2\Theta_{k-1}^2 \le h_k\ , \quad k = 1, 2, \ldots\,.$$

Bit counting lemma. The relative accuracy of these estimates is considered in the following lemma, the type of which will appear repeatedly in different context.

Lemma 2.4 *Assume that the just introduced Kantorovich estimates* $[h_k]$ *satisfy the relative accuracy requirement*

$$0 \le \frac{h_k - [h_k]}{[h_k]} \le \sigma < 1\,,\ k = 0, 1, \ldots\,.$$

Then

$$\Theta_{k+1} \le (1 + \sigma)\Theta_k^2\,,\ k = 0, 1, \ldots\,.$$

Proof. We collect the above relations to obtain

$$\Theta_{k+1} \leq \tfrac{1}{2}h_{k+1} \leq \tfrac{1}{2}(1+\sigma)[h_{k+1}] = (1+\sigma)\Theta_k^2 \,.$$

□

Restricted convergence monitor. With $\sigma \to 1$ we then end up with

$$\Theta_k \leq 2\Theta_{k-1}^2 \,, \qquad k = 0, 1, \dots \,,$$

which leads us to the requirement

$$\Theta_k \leq \tfrac{1}{2} \,, \quad k = 0, 1, \dots \,, \tag{2.12}$$

a convergence criterion more restrictive than (2.11) above. Otherwise we diagnose *divergence* of the ordinary Newton iteration.

Termination criterion. A desirable criterion to terminate the iteration would be

$$\|x^k - x^*\| \leq \mathrm{XTOL} \,, \tag{2.13}$$

with XTOL a user prescribed error tolerance. In view of (2.8) and with $h_k \to [h_k] = 2\Theta_{k-1}^2$ we will replace this condition by its cheaply computable substitute

$$\frac{\|\Delta x^k\|}{1 - \Theta_{k-1}^2} \leq \mathrm{XTOL} \,. \tag{2.14}$$

Note that XTOL can be chosen quite relaxed here, since $x^{k+1} = x^k + \Delta x^k$ is cheaply available with an accuracy of $O(\mathrm{XTOL}^2)$.

2.1.2 Simplified Newton method

Consider the simplified Newton iteration as introduced above:

$$F'(x^0)\overline{\Delta x}^k = -F(x^k) \,, \; x^{k+1} = x^k + \overline{\Delta x}^k \,, \; k = 0, 1, \dots \,. \tag{2.15}$$

Convergence analysis. We study the influence of the fixed initial Jacobian on the convergence behavior. The theorems to be derived are slight improvements of well-known theorems of J.M. Ortega and W.C. Rheinboldt—see [163].

Theorem 2.5 *Let $F : D \to \mathbb{R}^n$ be a continuously differentiable mapping with $D \subset \mathbb{R}^n$ open and convex. Let $x^0 \in D$ denote a given starting point so that $F'(x^0)$ is invertible. Assume the affine covariant Lipschitz condition*

$$\|F'(x^0)^{-1}\big(F'(x) - F'(x^0)\big)\| \leq \omega_0 \|x - x^0\| \tag{2.16}$$

for all $x \in D$. *Let*

$$h_0 := \omega_0 \|\overline{\Delta x}^0\| \leq \tfrac{1}{2} \tag{2.17}$$

and define

$$t^* = 1 - \sqrt{1 - 2h_0}\,, \;\; \rho = \frac{t^*}{\omega_0}\,.$$

Moreover, assume that $\bar{S}(x^0, \rho) \subset D$. *Then the simplified Newton iterates* (2.15) *remain in* $\bar{S}(x^0, \rho)$ *and converge to some* x^* *with* $F(x^*) = 0$. *The convergence rate can be estimated by*

$$\frac{\|x^{k+1} - x^k\|}{\|x^k - x^{k-1}\|} \leq \tfrac{1}{2}\big(t_k + t_{k-1}\big)\,, \;\; k = 1, 2, \ldots \tag{2.18}$$

and

$$\|x^k - x^*\| \leq \frac{t^* - t_k}{\omega_0}\,, \;\; k = 0, 1, \ldots \tag{2.19}$$

with $t_0 = 0$ *and*

$$t_{k+1} = h_0 + \tfrac{1}{2} t_k^2\,, \;\; k = 0, 1, \ldots\,.$$

Proof. We follow the line of the proofs in [163] and use (2.16) to obtain

$$\|x^{k+1} - x^k\| \leq \tfrac{1}{2}\omega_0 \|x^k - x^{k-1}\| \big(\|x^{k-1} - x^0\| + \|x^k - x^0\|\big)\,. \tag{2.20}$$

The result is slightly more complicated than for the ordinary Newton iteration. We therefore turn to a slightly more sophisticated proof technique by introducing the *majorants*

$$\omega_0 \|x^{k+1} - x^k\| \leq h_k\,, \;\; \omega_0 \|x^k - x^0\| \leq t_k$$

with initial values $t_0 = 0$, $h_0 \leq \frac{1}{2}$. Because of

$$\|x^{k+1} - x^0\| \leq \|x^k - x^0\| + \|x^{k+1} - x^k\|$$

and

$$\omega_0 \|x^{k+1} - x^k\| \leq \tfrac{1}{2} h_{k-1}(t_k + t_{k-1}) =: h_k$$

we select the two majorant equations

$$t_{k+1} = t_k + h_k\,, \;\; h_k = \tfrac{1}{2} h_{k-1}\big(t_k + t_{k-1}\big)\,,$$

which can be combined to a single equation of the form

$$t_{k+1} - t_k = (t_k - t_{k-1})\,\big(t_{k-1} + \tfrac{1}{2}(t_k - t_{k-1})\big) = \tfrac{1}{2}(t_k^2 - t_{k-1}^2)\,.$$

Rearrangement of this equation leads to

$$t_{k+1} - \tfrac{1}{2} t_k^2 = t_k - \tfrac{1}{2} t_{k-1}^2\,.$$

Since here the right hand side is just an index shift (downward) of the left hand side, we can apply the so-called *Ortega trick* to obtain

$$t_{k+1} - \tfrac{1}{2}t_k^2 = t_1 - \tfrac{1}{2}t_0^2 = h_0 \,,$$

which may be rewritten as the simplified Newton iteration

$$t_{k+1} - t_k = -\frac{g(t_k)}{g'(t_0)} = g(t_k)$$

for the scalar equation

$$g(t) = h_0 - t + \tfrac{1}{2}t^2 = 0 \,.$$

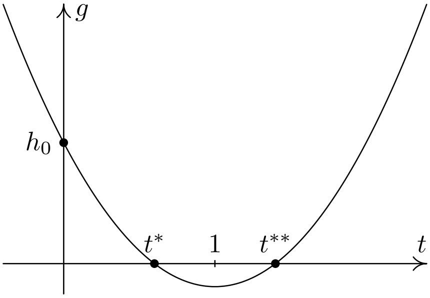

Fig. 2.1. Ortega trick: simplified Newton iteration.

As can be seen from Figure 2.1, the iteration starting at $t = 0$ will converge to the root t^*, which exists, if the above quadratic equation has two real roots. This implies the necessary condition $h_0 \leq 1/2$, which has been imposed above. Also from Figure 2.1 we immediately see that $g(t_{k+1}) < g(t_k)$, which is equivalent to $h_{k+1} < h_k$. Moreover with

$$t_k \leq t^* = 1 - \sqrt{1 - 2h_0} \,,$$

we immediately have

$$x^k \in \bar{S}(x^0, \rho) \subset D \,.$$

Hence, for the solution x^* we also get $x^* \in \bar{S}(x^0, \rho)$. As for the convergence rates, just observe that

$$\omega_0 \|x^k - x^*\| \leq \sum_{i=k}^{\infty} h_i = t^* - t_k$$

and use (2.20) to verify the remaining statements of the theorem. □

Convergence monitor. From Theorem 2.5 we derive that

$$\Theta_k = \frac{\|\overline{\Delta x}^{k+1}\|}{\|\overline{\Delta x}^{k}\|} \le \frac{h_{k+1}}{h_k} = \tfrac{1}{2}(t_{k+1} + t_k)\,.$$

With $t_0 = 0, t_1 = h_0$, the condition $h_0 \le 1/2$ induces the condition

$$\Theta_0 = \frac{\|\overline{\Delta x}^{1}\|}{\|\overline{\Delta x}^{0}\|} \le \tfrac{1}{2}h_0 \le \tfrac{1}{4}\,, \tag{2.21}$$

which characterizes the local convergence domain of the simplified Newton method. In comparison with $\Theta_0 < 1$ for the ordinary Newton method, where a new Jacobian is used at each step, this is a clear reduction. The above result also shows that the convergence rate may slow down to

$$\Theta_k < t^* = 1 - \sqrt{1 - 2h_0}\,.$$

We may replace the theoretical quantity t^* by its computationally available bounds

$$[t^*] = 1 - \sqrt{1 - 4\Theta_0} \le 1 - \sqrt{1 - 2h_0} = t^* \le 1\,.$$

Then *divergence* of the simplified Newton iteration will be defined to occur when $\Theta_k \ge [t^*]$.

Termination criterion. From (2.19) we may derive the upper bound

$$\|x^k - x^*\| \le \frac{t^* - t_k}{\omega_0}\,.$$

This line is just a different form of the repeated triangle inequality used in the proof so that

$$\|x^k - x^*\| \le \sum_{j=k}^{\infty} \|\overline{\Delta x}^{j}\|\,.$$

This gives rise to the upper bound

$$\|x^k - x^*\| \le \|\overline{\Delta x}^{k}\|\,(1 + \Theta_k + \Theta_{k+1}\Theta_k + \ldots) \le \frac{\|\overline{\Delta x}^{k}\|}{1 - t^*}\,.$$

Upon insertion of the estimate $[t^*] \le t^*$ from above, we are led to the *approximate* termination criterion

$$\frac{\|\overline{\Delta x}^{k}\|}{\sqrt{1 - 4\Theta_0}} \le \mathrm{XTOL}\,,$$

where XTOL is the user prescribed final error tolerance. Of course, the application of such a criterion will require to start with some $\Theta_0 < \frac{1}{4}$.

2.1.3 Newton-like methods

Consider a rather general Newton-like iteration of the form

$$M(x^k)\delta x^k = -F(x^k)\,,\ x^{k+1} = x^k + \delta x^k\,, \qquad k = 0, 1, \dots\,. \tag{2.22}$$

Convergence analysis. From the basic construction idea, such an iteration will converge, if $M(x)$ is a 'sufficiently accurate' approximation of $F'(x)$. The question will be how to measure the approximation quality and to quantify the vague term 'sufficiently accurate'.

Theorem 2.6 *Let $F : D \to \mathbb{R}^n$ be a continuously differentiable mapping with $D \subset \mathbb{R}^n$ open and convex. Let M denote an approximation of F'. Assume that one can find a starting point $x^0 \in D$ with $M(x^0)$ invertible and constants α, $\overline{\omega}_0$, δ_0, δ_1, $\delta_2 \geq 0$ such that for all $x, y \in D$*

$$\big\|M(x^0)^{-1}F(x^0)\big\| \leq \alpha\,,$$

$$\big\|M(x^0)^{-1}\big(F'(y) - F'(x)\big)\big\| \leq \overline{\omega}_0\|y - x\|\,,$$

$$\big\|M(x^0)^{-1}\big(F'(x) - M(x)\big)\big\| \leq \delta_0 + \delta_1\|x - x^0\|\,,$$

$$\big\|M(x^0)^{-1}\big(M(x) - M(x^0)\big)\big\| \leq \delta_2\|x - x^0\|\,,$$

$$\delta_0 < 1,\ \ \sigma := \max(\overline{\omega}_0, \delta_1 + \delta_2),\ \ h := \frac{2\alpha\sigma}{(1 - \delta_0)^2} \leq 1\,, \tag{2.23}$$

$$\overline{S}(x^0, \rho) \subset D \ \ \textit{with} \ \ \rho := \frac{2\alpha}{1 - \delta_0} \ / \ \Big(1 + \sqrt{1 - h}\Big)\,.$$

Then the sequence $\{x^k\}$ generated from the Newton-like iteration (2.22) *is well-defined, remains in $\overline{S}(x^0, \rho)$ and converges to a solution point x^* with $F(x^*) = 0$. With the notation*

$$\overline{h} := \frac{\overline{\omega}_0}{\sigma}h,\ \ \rho_\pm = \frac{2\alpha}{1 - \delta_0} \ / \ \big(1 \mp \sqrt{1 - \overline{h}}\,\big)$$

the solution $x^* \in \overline{S}(x^0, \rho_-)$ is unique in

$$\overline{S}(x^0, \rho) \cup \big(D \cap S(x^0, \rho_+)\big)\,.$$

Proof. For the usual induction proof, the following majorants are convenient

$$\begin{array}{rcllcl} \overline{\omega}_0\|\delta x^k\| & \leq & h_k\,, & h_0 & := & \alpha\overline{\omega}_0\,, \\ \overline{\omega}_0\|x^k - x^0\| & \leq & t_k\,, & t_0 & := & 0\,, \end{array}$$

together with

$$t_{k+1} = t_k + h_k\,. \tag{2.24}$$

Proceeding as in the proofs of the preceding theorems, one obtains

$$\begin{aligned}&\|x^{k+1}-x^k\| = \|M(x^k)^{-1}F(x^k)\|\\ &\quad = \left\|M(x^k)^{-1}\left[F(x^k)-\left(F(x^{k-1})+M(x^{k-1})(x^k-x^{k-1})\right)\right]\right\|\\ &\quad \le \left\|M(x^k)^{-1}\left[F(x^k)-F(x^{k-1})-F'(x^{k-1})(x^k-x^{k-1})\right]\right\|\\ &\qquad + \left\|M(x^k)^{-1}\left[F'(x^{k-1})-M(x^{k-1})\right](x^k-x^{k-1})\right\| \,.\end{aligned}$$

The perturbation lemma yields:

$$\|M(x^k)^{-1}M(x^0)\| \le 1 \;/\; (1-\delta_2\|x^k-x^0\|)\,.$$

Combining these intermediate results and using $\overline{\delta}_i := \delta_i/\overline{\omega}_0$, $i=1,2$, then supplies

$$\overline{\omega}_0\|x^{k+1}-x^k\| \le \frac{1}{1-\overline{\delta}_2 t_k}\left[\tfrac{1}{2}h_{k-1}^2+\left(\delta_0+\overline{\delta}_1 t_{k-1}\right)h_{k-1}\right] \,.$$

Thus one ends up with the second majorant equation

$$h_k = \left[\left(\delta_0+\overline{\delta}_1 t_{k-1}\right)h_{k-1}+\tfrac{1}{2}h_{k-1}^2\right] \;/\; \left(1-\overline{\delta}_2 t_k\right) \,.$$

Reformulation in view of a possible application of the Ortega technique leads to

$$\begin{aligned}&\left(1-\overline{\delta}_2 t_k\right)h_k - \left(1-\overline{\delta}_2 t_{k-1}\right)h_{k-1}\\ &\quad = \tfrac{1}{2}\left(t_k^2-t_{k-1}^2\right)-(1-\delta_0)\left(t_k-t_{k-1}\right)\\ &\qquad + \left(\overline{\delta}_1+\overline{\delta}_2-1\right)\left(t_k t_{k-1}-t_{k-1}^2\right) \,.\end{aligned} \tag{2.25}$$

Obviously, this technique is only applicable, if one requires that

$$\overline{\delta}_1+\overline{\delta}_2 = 1 \text{ , i.e., } \delta_1+\delta_2=\overline{\omega}_0 \,,$$

which will not be the case in general. However, by defining σ as in assumption (2.23) and redefining

$$\begin{aligned}\sigma\|\delta x^k\| &\le h_k\,, & h_0 &:= \alpha\sigma\,,\\ \sigma\|x^k-x^0\| &\le t_k\,, & t_0 &:= 0\,,\\ \overline{\delta}_i &:= \delta_i/\sigma\,, & i &= 1,2\end{aligned}$$

the disturbing term in (2.25) will vanish. Insertion of (2.24) then gives

$$\left(1-\overline{\delta}_2 t_k\right)\left(t_{k+1}-t_k\right)+(1-\delta_0)\,t_k-\tfrac{1}{2}t_k^2 = t_1 = \alpha\sigma\,,$$

which can be rewritten in the form

$$t_{k+1} - t_k = \frac{h_0 - (1-\delta_0)t_k + \frac{1}{2}t_k^2}{1 - \overline{\delta}_2 t_k}\,.$$

This iteration can be interpreted as a Newton-like iteration in $\mathbb{R}^1$ for the solution of

$$g(t) := h_0 - (1-\delta_0)t + \tfrac{1}{2}t^2 = 0\,.$$

The associated two roots

$$\begin{aligned} t^* &= (1-\delta_0)\left(1 - \sqrt{1 - \frac{2\alpha\sigma}{(1-\delta_0)^2}}\right)\,, \\ t^{**} &= (1-\delta_0)\left(1 + \sqrt{1 - \frac{2\alpha\sigma}{(1-\delta_0)^2}}\right) \end{aligned}$$

are real if

$$\frac{2\alpha\sigma}{(1-\delta_0)^2} \le 1\,.$$

This is just assumption (2.23). The remaining part of the proof essentially follows the lines of the proof of Theorem 2.5 and is therefore omitted here. □

The above theorem does not supply any direct advice towards algorithmic realization. In practical applications, however, additional structure on the approximations $M(x)$ will be given—often as a dependence on an additional parameter, which can be manipulated in such a way that convergence criteria can be met. A typical version of Newton-like methods is the deliberate dropping of 'weak couplings' in the derivative, which can be neglected on the basis of insight into the specific underlying problem. In finite dimensions, deliberate 'sparsing' can be used, which means dropping of 'small' entries in a large Jacobian matrix; this technique works efficiently, if the vague term 'small' can be made sufficiently precise from the application context. Needless to say that 'sparsing' nicely goes with sparse matrix techniques.

2.1.4 Broyden's 'good' rank-1 updates

In order to derive an *error oriented quasi-Newton method*, we start by rewriting the secant condition (1.17) strictly in *affine covariant* terms of quantities in the domain space of F. This leads to

$$E_k(J)\,\delta x_k = \overline{\delta x}_{k+1} = -J_k^{-1}F_{k+1}$$

in terms of the affine covariant update change matrix

$$E_k(J) := I - J_k^{-1}J\,.$$

Any Jacobian rank-1 update of the kind

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

with v some vector in the domain space of F will both satisfy the secant condition and exhibit the here desired affine covariance property. The update with $v = \delta x_k$ is known in the literature as 'good Broyden update' [40].

Auxiliary results. The following theorem will collect a bunch of useful results for a single iterative step of the thus defined quasi-Newton method.

Theorem 2.7 *In the notation just introduced, let*

$$J_{k+1} = J_k \left(I - \frac{\overline{\delta x}_{k+1} \delta x_k^T}{\|\delta x_k\|_2^2} \right) \tag{2.26}$$

denote an affine covariant Jacobian rank-1 update and assume the local contraction condition

$$\Theta_k = \frac{\|\overline{\delta x}_{k+1}\|_2}{\|\delta x_k\|_2} < \tfrac{1}{2} .$$

Then:

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

$$\begin{aligned} \|E_k(J_{k+1})\|_2 &\leq \|E_k(J)\|_2 , \; \forall \; J \in \mathcal{S}_k , \\ \|E_k(J_{k+1})\|_2 &\leq \Theta_k . \end{aligned}$$

(II) *The update matrix J_{k+1} is nonsingular whenever J_k is nonsingular, and its inverse can be represented in the form*

$$J_{k+1}^{-1} = \left(I + \frac{\overline{\delta x}_{k+1} \delta x_k^T}{(1 - \alpha_{k+1}) \|\delta x_k\|_2^2} \right) J_k^{-1} \tag{2.27}$$

with

$$\alpha_{k+1} = \frac{\delta x_k^T \overline{\delta x}_{k+1}}{\|\delta x_k\|_2^2} < \tfrac{1}{2} .$$

(III) *The next quasi-Newton correction is*

$$\delta x_{k+1} = -J_{k+1}^{-1} F_{k+1} = \frac{\overline{\delta x}_{k+1}}{1 - \alpha_{k+1}} .$$

(IV) *Iterative contraction in terms of quasi-Newton corrections shows up as*

$$\frac{\|\delta x_{k+1}\|_2}{\|\delta x_k\|_2} = \frac{\Theta_k}{1 - \alpha_{k+1}} < 1 .$$

Proof. For the rank-1 update we directly have

$$E_k(J_{k+1}) = \frac{\overline{\delta x}_{k+1}\delta x_k^T}{\|\delta x_k\|_2^2} \Longrightarrow \|E_k(J_{k+1})\|_2 \leq \Theta_k\,.$$

As for the least change update property, we obtain

$$\|E_k(J_{k+1})\|_2 = \left\|\frac{\overline{\delta x}_{k+1}\delta x_k^T}{\|\delta x_k\|_2^2}\right\|_2 = \left\|E_k(J)\frac{\delta x_k \delta x_k^T}{\|\delta x_k\|_2^2}\right\|_2 \leq \|E_k(J)\|_2\,,$$

which confirms statement I. By application of the Sherman-Morrison formula (see, for instance, the book of A.S. Householder [121]), we directly verify the statements II and III. In order to show IV, we apply the Cauchy-Schwarz inequality to see that

$$|\alpha_{k+1}| \leq \Theta_k,$$

which, for $\Theta_k < 1/2$, implies

$$\frac{\|\delta x_{k+1}\|}{\|\delta x_k\|} = \frac{\Theta_k}{1-\alpha_{k+1}} \leq \frac{\Theta_k}{1-\Theta_k} < 1\,.$$

□

Algorithmic realization. The result (2.27) may be rewritten as

$$J_{k+1}^{-1} = \left(I + \frac{\delta x_{k+1}\delta x_k^T}{\|\delta x_k\|_2^2}\right) J_k^{-1}\,.$$

This recursion cannot be used directly for the computation of δx_{k+1}. However, the product representation

$$J_k^{-1} = \left(I + \frac{\delta x_k \delta x_{k-1}^T}{\|\delta x_{k-1}\|_2^2}\right) \cdot \dots \cdot \left(I + \frac{\delta x_1 \delta x_0^T}{\|\delta x_0\|_2^2}\right) J_0^{-1}\,.$$

can be applied up to the correction δx_k. This consideration leads to a rather economic *recursive 'good' Broyden algorithm*, which has been used for quite a while in the public domain code NLEQ1 [161]. It essentially requires the $k_{\max}+1$ quasi-Newton corrections $\delta x_0, \dots, \delta x_k$ as extra array storage.

Discrete norms for differential equations. Inner products $\langle u, v\rangle$ other than the Euclidean inner product $u^T v$ may be used in view of the underlying problem—such as (discrete) Sobolev inner products for discretized differential equations. By all means, *scaling in the domain space* of F should be carefully considered. This means that any corrections δx arising in the above inner products should actually be implemented as $D^{-1}\delta x$ with appropriate diagonal scaling matrix D. If D is chosen in agreement with a *relative error* concept, then in this way *scaling invariance* of the algorithm can be assured.

Condition number monitor. Recursive implementations based on the above rank-1 factorization have often been outruled with the argument that some hidden ill-conditioning in the arising Jacobian updates might occur. In order to derive a some monitor, we may use

$$\mathrm{cond}_2(J_{k+1}) \leq \mathrm{cond}_2\left(I + \frac{\delta x_{k+1}\delta x_k^T}{\|\delta x_k\|_2^2}\right)\mathrm{cond}_2(J_k)\ .$$

In this context, the following technical lemma may be helpful.

Lemma 2.8 *Given a rank-1 matrix*

$$A = I - \frac{uv^T}{v^T v} \ \ \textit{with}\ \Theta := \frac{\|u\|_2}{\|v\|_2} < 1\,,$$

its condition number can be bounded as

$$\mathrm{cond}_2(A) \leq \frac{1+\Theta}{1-\Theta}\,.$$

Proof. We just use the two bounds

$$\|A\| \leq 1 + \left\|\frac{uv^T}{v^T v}\right\| \leq 1+\Theta, \quad \|A^{-1}\| \leq \left(1 - \left\|\frac{uv^T}{v^T v}\right\|\right)^{-1} \leq (1-\Theta)^{-1}$$

and insert into the definition $\mathrm{cond}_2(A) = \|A\|\,\|A^{-1}\|$. □

With this result and $\Theta_k < 1/2$ we are certainly able to assure that

$$\mathrm{cond}_2(J_{k+1}) \leq \frac{1+\Theta_k}{1-\Theta_k}\mathrm{cond}_2(J_k) < 3\ \ \mathrm{cond}_2(J_k)\,.$$

Convergence monitor. In accordance with the above theoretical results, we impose the condition $\Theta_k < 1/2$ throughout the whole iteration. Note that this is an extension of the local convergence domain compared with the simplified Newton method where $\Theta_0 \leq 1/4$ has to be required. With these preparations we are now ready to state the 'good Broyden algorithm' QNERR (for ERRor oriented Quasi-Newton method) in the usual informal manner.

Algorithm QNERR.

```
For given  $x^0$:  $F_0 = F(x^0)$                evaluation and store
                   $J_0\delta x_0 = -F_0$          linear system solve
                   $\sigma_0 = \|\delta x_0\|_2^2$   store $\delta x_0, \sigma_0$
```

```
For $k = 0, \ \dots, \ k_{\max}$:
    I.    $x^{k+1} = x^k + \delta x_k$                          new iterate
          $F_{k+1} = F(x^{k+1})$                                evaluation
          $J_0 v = -F_{k+1}$                                     linear system solve

    II.   If $k > 0$: for $i = 1, \dots, k$
                  $\overline{\alpha} := \frac{v^T \delta x_{i-1}}{\sigma_{i-1}}$ ,
                  $v := v + \overline{\alpha} \delta x_i$

    III.  Compute
          $\alpha_{k+1} := \frac{v^T \delta x_k}{\sigma_k}$ , $\Theta_k = \left( \frac{v^T v}{\sigma_k} \right)^{1/2}$        store
          If $\Theta_k > \frac{1}{2}$: stop, no convergence
          IV.   $\delta x_{k+1} = \frac{v}{1 - \alpha_{k+1}}$ ,                 store
                $\sigma_{k+1} = \|\delta x_{k+1}\|_2^2$                      store
                If $\sqrt{\sigma_{k+1}} \leq \mathrm{XTOL}$:
                solution $x^* = x^{k+1} + \delta x_{k+1}$
Else:     no convergence within $k_{\max}$ iterations.
```

Convergence analysis. The above Theorem 2.7 does not give conditions, under which the contraction condition $\Theta_k < 1/2$ is assured *throughout the whole iteration.* This will be the topic of the next theorem.

Theorem 2.9 *For $F : D \longrightarrow \mathbb{R}^n$ be a continuously differentiable mapping with D open and convex. Let $x^* \in D$ denote a unique solution point of F with $F'(x^*)$ nonsingular. Assume that the following affine covariant Lipschitz condition holds:*

$$\|F'(x^*)^{-1}\big(F'(x) - F'(x^*)\big)v\| \leq \omega \|x - x^*\| \cdot \|v\|$$

for $x, \ x + v \in D$ and $0 \leq \omega < \infty$. Consider the quasi-Newton iteration as defined in Theorem 2.7. For some $\overline{\Theta}$ in the range $0 < \overline{\Theta} < 1$ assume that:

(I) *the initial approximate Jacobian J_0 satisfies*

$$\delta_0 := \big\|F'(x^*)^{-1}\big(J_0 - F'(x^0)\big)\big\| < \overline{\Theta}/(1 + \overline{\Theta}) \, , \tag{2.28}$$

(II) *the initial guess* x^0 *satisfies*

$$t_0 := \omega\|x^0 - x^*\| \leq \frac{1-\overline{\Theta}}{2-\overline{\Theta}}\left(\frac{\overline{\Theta}}{1+\overline{\Theta}} - \delta_0\right). \tag{2.29}$$

Then the quasi-Newton iterates $\{x^k\}$ *converge to* x^* *in terms of errors as*

$$\|x^{k+1} - x^*\| < \overline{\Theta}\|x^k - x^*\|, \tag{2.30}$$

or, in terms of corrections as

$$\|\delta x^{k+1}\| \leq \overline{\Theta}\|\delta x^k\|. \tag{2.31}$$

The convergence is superlinear with

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

As for the Jacobian rank-1 updates, the 'bounded deterioration property' holds in the form

$$\|E_k\| := \|F'(x^*)^{-1}J_k - I\| \leq \frac{\overline{\Theta}}{1+\overline{\Theta}} < \tfrac{1}{2} \tag{2.32}$$

together with the asymptotic property

$$\lim_{k\to\infty} \frac{\|E_k\delta x_k\|}{\|\delta x_k\|} = 0. \tag{2.33}$$

Proof. Let $\|\cdot\|$ be $\|\cdot\|_2$ throughout. For ease of writing we characterize the Jacobian update approximation by

$$\eta_k = \frac{\|E_k\delta x_k\|}{\|\delta x_k\|}, \overline{\eta}_k = \|E_k\| = \|E_k^T\| = \max_{v\neq 0}\frac{\|E_k^{(T)}v\|}{\|v\|}.$$

By definition, $\eta_k \leq \overline{\eta}_k$. For the convergence analysis we introduce

$$t_k = \omega\|e_k\|, \; e_k := x^k - x^*.$$

As usual [57], the proof is performed in two basic steps: first *linear* convergence, then *superlinear* convergence.

I. To begin with, exploit the *Lipschitz condition* in the form

$$\begin{aligned}\|F'(x^*)^{-1}F(x^{k+1})\| &\leq \int_{s=0}^{1} \left\|F'(x^*)^{-1}\big(F'(x^k + s\delta x_k) - F'(x^*)\big)\delta x_k\right\| ds \\ &\quad + \|E_k\delta x_k\| \\ &\leq \big(\tfrac{1}{2}(t_{k+1} + t_k) + \eta_k\big)\|\delta x_k\|.\end{aligned}$$

Under the assumption $\eta_{k+1} < 1$ we may estimate

$$\|F'(x^*)^{-1}F(x^{k+1})\| = \|(I + E_{k+1})\delta x_{k+1}\| \geq (1 - \eta_{k+1})\|\delta x_{k+1}\|$$

so that

$$\frac{\|\delta x_{k+1}\|}{\|\delta x_k\|} \leq \frac{\eta_k + \bar{t}_k}{1 - \eta_{k+1}}\,, \quad \text{where} \quad \bar{t}_k := \tfrac{1}{2}(t_k + t_{k+1})\,. \tag{2.34}$$

As for the iterative errors e_k, we may derive the relation

$$\begin{aligned} e_{k+1} &= e_k - J_k^{-1}F(x^k) \\ &= (I + E_k)^{-1}\left(E_k e_k - F'(x^*)^{-1}\int\limits_{s=0}^{1}\big(F'(x^* + se_k) - F'(x^*)\big)e_k ds\right), \end{aligned}$$

from which we obtain the estimate (let $\overline{\eta}_k < 1$)

$$t_{k+1} \leq \frac{\overline{\eta}_k + \frac{1}{2}t_k}{1 - \overline{\eta}_k}t_k\,. \tag{2.35}$$

Upon comparing the right hand upper bounds in (2.35) and (2.34) we are led to define the majorant

$$\overline{\Theta} := \frac{\overline{\eta}_k + \bar{t}_k}{1 - \overline{\eta}_k}\,, \tag{2.36}$$

which implies that

$$t_{k+1} < \overline{\Theta} t_k\,. \tag{2.37}$$

II. Next, we study the *approximation properties* of the Jacobian updates. With E_k as defined, the above rank-1 update may be rewritten in the form

$$E_{k+1} = E_k + F'(x^*)^{-1}\frac{F_{k+1}\delta x_k^T}{\|\delta x_k\|_2^2}\,.$$

If we insert

$$F'(x^*)^{-1}F_{k+1} = (D_{k+1} - E_k)\delta x_k\,,$$

wherein

$$D_{k+1} := F'(x^*)^{-1}\int\limits_{s=0}^{1}\big(F'(x^k + s\delta x_k) - F'(x^*)\big)ds$$

and introduce the orthogonal projections

$$Q_k^{\perp} = I - Q_k = \frac{\delta x_k \delta x_k^T}{\|\delta x_k\|^2}\,,$$

then we arrive at the decomposition

$$E_{k+1} = E_k Q_k + D_{k+1}Q_k^{\perp}$$

and its transpose ($v \neq 0$ arbitrary)

$$E_{k+1}^T v = Q_k E_k^T v + Q_k^{\perp} D_{k+1}^T v \,. \tag{2.38}$$

Note that

$$\frac{\|E_{k+1}\delta x_k\|}{\|\delta x_k\|} = \frac{\|D_{k+1}\delta x_k\|}{\|\delta x_k\|} \leq \bar{t}_k \,. \tag{2.39}$$

III. In order to prove *linear* convergence, equation (2.38) is used for the quite rough estimate

$$\overline{\eta}_{k+1} = \max_{v \neq 0} \frac{\|E_{k+1}^T v\|}{\|v\|} \leq \max_{v \neq 0} \frac{\|E_k^T v\|}{\|v\|} + \max_{v \neq 0} \frac{|\langle D_{k+1}\delta x_k, v\rangle|}{\|\delta x_k\|\|v\|} \leq \overline{\eta}_k + \bar{t}_k \,. \tag{2.40}$$

Assume now that we have *uniform* upper bounds

$$\overline{\Theta} \leq \overline{\Theta} < 1 \,, \; \overline{\eta}_k \leq \overline{\eta} < 1 \,.$$

Then (2.37) can be replaced by

$$t_{k+1} < \overline{\Theta} t_k < t_k$$

and (2.36) leads to the natural definition

$$\overline{\Theta} \leq \frac{\overline{\eta} + \bar{t}_0}{1 - \overline{\eta}} =: \overline{\Theta} \,. \tag{2.41}$$

As for the definition of $\overline{\eta}$, we apply (2.40) to obtain

$$\overline{\eta}_{k+1} < \overline{\eta}_0 + \sum_{l=0}^{k} \bar{t}_l < \overline{\eta}_0 + \frac{\bar{t}_0}{1 - \overline{\Theta}} =: \overline{\eta} \,. \tag{2.42}$$

Insertion of $\overline{\eta}$ into (2.41) then eventually yields after some calculation:

$$t_0 \leq (1 - \overline{\Theta}) \left(\frac{\overline{\Theta}}{1 + \overline{\Theta}} - \overline{\eta}_0 \right) \,, \tag{2.43}$$

which obviously requires

$$\overline{\eta}_0 < \frac{\overline{\Theta}}{1 + \overline{\Theta}} < \tfrac{1}{2} \;\text{ for }\; \overline{\Theta} < 1 \,.$$

Observe now that by mere triangle inequality, with δ_0 as defined in (2.28), we have $\overline{\eta}_0 \leq t_0 + \delta_0$. Therefore, the assumption (2.43) can finally be replaced by the above two assumptions (2.28) and (2.29). Once such a $\overline{\Theta} < 1$ exists, we have (2.30) directly from $t_{k+1} < \overline{\Theta} t_k$ and (2.31) from inserting $\overline{\eta}$ into (2.34). The bounded deterioration property (2.32) follows by construction and insertion of (2.29) into (2.42).

IV. In order to show *superlinear* convergence, we use (2.38) in a more subtle manner. In terms of the *Euclidean* inner product $\langle \cdot, \cdot \rangle$, some short calculation supplies the equation

$$\|E_{k+1}^T v\|^2 = \|E_k^T v\|^2 - \frac{\langle E_k \delta x_k,\, v\rangle^2}{\|\delta x_k\|^2} + \frac{\langle D_{k+1}\delta x_k,\, v\rangle^2}{\|\delta x_k\|^2}\,.$$

Summing over the indices k, we arrive at

$$\sum_{k=0}^{l} \frac{\langle E_k \delta x_k, v\rangle^2}{\|v\|^2\|\delta x_k\|^2} = \frac{\|E_0^T v\|^2}{\|v\|^2} - \frac{\|E_{l+1}^T v\|^2}{\|v\|^2} + \sum_{k=0}^{l} \frac{\langle D_{k+1}\delta x_k, v\rangle^2}{\|\delta x_k\|^2\|v\|^2}\,.$$

Upon dropping the negative right hand term, letting $l \to \infty$, and using (2.39) with $\overline{t}_{k+1} < \overline{\Theta} \cdot \overline{t}_k$, we end up with the estimate

$$\sum_{k=0}^{\infty} \frac{\langle E_k \delta x_k,\, v\rangle^2}{\|v\|^2\|\delta x_k\|^2} \le \overline{\eta}_0^2 + \tfrac{1}{2}\frac{1+\overline{\Theta}}{1-\overline{\Theta}} t_0^2\,.$$

Since the right hand side is bounded, we immediately conclude that

$$\lim_{k\to\infty} \frac{\langle E_k \delta x_k,\, v\rangle^2}{\|\delta x_k\|^2\|v\|^2} = 0 \quad \forall\, v \in \mathbb{R}^n\,.$$

As a consequence, with

$$\xi_k := \frac{\delta x_k}{\|\delta x_k\|}\,,$$

we must have

$$\lim_{k\to\infty} E_k \xi_k = 0$$

from which statement (2.33) follows. Finally, with (2.34), we have proved superlinear convergence. □

Bibliographical Note. Quasi-Newton methods are described, e.g., in the classical optimization book [57] by J.E. Dennis and R.B. Schnabel or, more recently, in the textbook [132] by C.T. Kelley. These methods essentially started with the pioneering paper [40] by C.G. Broyden. For quite a time, the convergence of the 'good' Broyden method was not at all clear. A breakthrough in its convergence analysis came by the paper [41] of C.G. Broyden, J.E. Dennis, and J.J. Moré, where local and superlinear convergence has been shown on the basis of condition (2.33), the meanwhile so-called *Dennis-Moré condition* (see [55]). To the most part, the present section is an affine covariant reformulation of well-known material spread over a huge literature—see, e.g., the original papers [56] by J.E. Dennis and R.B. Schnabel or [58] by J.E. Dennis and H.F. Walker.

The above quasi–Newton algorithm is realized within the earlier code `NLEQ1` and its update `NLEQ-ERR`.

2.1.5 Inexact Newton-ERR methods

Inexact Newton methods consist of a combination of an outer iteration, the Newton iteration, and an inner iteration such that (dropping the inner iteration index i)

$$F'(x^k)(\delta x^k - \Delta x^k) = r^k \ , \quad x^{k+1} = x^k + \delta x^k \ , \quad k = 0, 1, \dots \, .$$

Here the inner residual r^k gives rise to the difference between the exact Newton correction Δx^k and the inexact Newton correction δx^k. Among the possible inner iterative solvers we will concentrate on those that reduce the *Euclidean error norms* $\|\delta x^k - \Delta x^k\|$, which leads us to CGNE (compare Section 1.4.3) and to GBIT (compare Section 1.4.4). In both cases, the perturbation will be measured by the relative difference between the exact Newton correction Δx^k and the inexact Newton correction δx^k via

$$\delta_k = \frac{\|\delta x^k - \Delta x^k\|}{\|\delta x^k\|} \, , \ k = 0, 1, \dots \, . \tag{2.44}$$

As a guiding principle for convergence, we will focus on contraction in terms of the (not actually computed) *exact* Newton corrections

$$\Theta_k = \frac{\|\Delta x^{k+1}\|}{\|\Delta x^k\|} \, ,$$

subject to the perturbation coming from the truncation of the inner iteration.

Convergence analysis—CGNE. First we work out details for the error *minimizing* case, exemplified by CGNE specifying the norm $\|\cdot\|$ to be the Euclidean norm $\|\cdot\|_2$. Upon recalling (1.28), the starting value $\delta x_0^k = 0$ for the CGNE iteration implies that

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

Moreover, from (1.29) and (1.30) we conclude that δ_k is monotonically decreasing in the course of the inner iteration so that eventually any threshold condition of the type $\delta_k \leq \bar{\delta}$ can be met. With this preparation, we are now ready to state our convergence result.

Theorem 2.10 *Let $F : D \longrightarrow \mathbb{R}^n$ be a continuously differentiable mapping with $D \subset \mathbb{R}^n$ open, convex, and sufficiently large. Suppose that $F'(x)$ is invertible for each $x \in D$. Assume that the following* affine covariant *Lipschitz condition holds:*

$$\|F'(z)^{-1}\big(F'(y) - F(x)\big)v\| \leq \omega \|y - x\| \cdot \|v\|$$

$$\text{for collinear } \ x, \, y, \, z \in D \, .$$

Let $x^0 \in D$ denote a given starting point for a Newton-CGNE iteration. At an iterate x^k, let δ_k as defined in (2.44) denote the relative error of the inexact Newton correction δx^k. Let the inner CGNE *iteration be started with $\delta x_0^k = 0$, which gives rise to the following relations between the Kantorovich quantities*

$$h_k := \omega \|\Delta x^k\| \quad \text{and} \quad h_k^\delta := \omega \|\delta x^k\| = \frac{h_k}{\sqrt{1+\delta_k^2}}\,.$$

Let $x^* \in D$ be the unique solution point.

I. Linear convergence mode. *Assume that an initial guess x^0 has been chosen such that*

$$h_0 < 2\overline{\Theta} < 2$$

for some $\overline{\Theta} < 1$. Let $\delta_{k+1} \geq \delta_k$ be realized throughout the inexact Newton iteration and control the inner iteration such that

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

which assures that

$$\delta_k \leq \frac{\overline{\Theta}}{\sqrt{1-\overline{\Theta}^2}}\,. \tag{2.45}$$

Then this implies the exact monotonicity

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

and the inexact monotonicity

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

The iterates $\{x^k\}$ remain in $\overline{S}(x^0, \rho)$ with $\rho = \|\delta x^0\|/(1-\overline{\Theta})$ and converge at least linearly to x^*.

II. Quadratic convergence mode. *For some $\rho > 0$, let the initial guess x^0 satisfy*

$$h_0 < \frac{2}{1+\rho} \tag{2.46}$$

and control the inner iteration such that

$$\delta_k \leq \frac{\rho}{2}\frac{h_k^\delta}{1+h_k^\delta}\,, \tag{2.47}$$

which requires that

$$\rho > \frac{3\delta_0}{1-\delta_0} \tag{2.48}$$

be chosen. Then the inexact Newton iterates remain in $\overline{S}(x^0, \overline{\rho})$ with

$$\overline{\rho} = \|\delta x^0\| / \left(1 - \frac{1+\rho}{2} h_0\right)$$

and converge quadratically to x^* with

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

and

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

Proof. First we show that

$$\|\Delta x^{k+1}\| \leq \int\limits_{t=0}^{1} \|F'(x^{k+1})^{-1}\big(F'(x^k + t\delta x^k) - F'(x^k)\big)\delta x^k\| dt + \|F'(x^{k+1})^{-1} r^k\| \, .$$

For the first term we just apply the Lipschitz condition in standard form. For the second term we may use the same condition plus the triangle inequality to obtain

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

With definition (2.44), this gives

$$\frac{\|\Delta x^{k+1}\|}{\|\delta x^k\|} \leq \tfrac{1}{2} h_k^\delta + \delta_k (1 + h_k^\delta) \, . \tag{2.49}$$

With $h_k^\delta = h_k / \sqrt{1 + \delta_k^2}$ we then arrive at

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

In order to prove *linear* convergence, we might require $\vartheta(h_k, \delta_k) = \overline{\Theta} < 1$, which implies that δ_k monotonically *increases* as h_k monotonically decreases—which would automatically lead to $\delta_{k+1} \geq \delta_k$ when $h_{k+1} \leq h_k$. However, since strict equality cannot be realized within `CGNE`, we have to assume the two separate inequalities $\vartheta \leq \overline{\Theta}$ and $\delta_{k+1} \geq \delta_k$, as done in the theorem. Note that a necessary condition for $\vartheta(h_k, \delta_k) \leq \overline{\Theta}$ with some $\delta_k > 0$ is that it holds at least for $\delta_k = 0$, which yields $h_0 < 2\overline{\Theta}$, the assumption made in the theorem.

As for the contraction in terms of the inexact Newton corrections, we then obtain

$$\frac{\|\delta x^{k+1}\|}{\|\delta x^k\|} = \sqrt{\frac{1+\delta_k^2}{1+\delta_{k+1}^2}}\,\frac{\|\Delta x^{k+1}\|}{\|\Delta x^k\|} \le \sqrt{\frac{1+\delta_k^2}{1+\delta_{k+1}^2}}\,\overline{\Theta} \le \overline{\Theta}\,.$$

Usual linear convergence results then imply that $\{x^k\}$ remains in $\overline{S}(x^0,\rho)$ with $\rho = \|\delta x^0\|/(1-\overline{\Theta})$, if only $\overline{S}(x^0,\rho) \subset D$, which we assumed by D to be 'sufficiently large'. Asymptotically we thus assure that $\vartheta(0,\delta_k) \le \overline{\Theta}$, which is equivalent to (2.45).

For the *quadratic* convergence case we require that the first term in $\vartheta(h_k,\delta_k)$ originating from the outer iteration exceeds the second term, which brings us to (2.47). Note that now $h_{k+1} \le h_k$ implies $\delta_{k+1} \le \delta_k$ and $h_k \to 0$ also $\delta_k \to 0$—a behavior that differs from the linear convergence case. Insertion of (2.47) into $\vartheta(h_k,\delta_k)$ then directly leads to

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

and to

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

Upon applying the usual quadratic convergence results, we have to require the sufficient condition

$$\frac{1+\rho}{2} h_0^\delta \le \frac{1+\rho}{2} h_0 < 1$$

and then, assuming that D is 'sufficiently large', obtain convergence within the ball

$$\overline{S}(x^0,\overline{\rho})\,,\ \overline{\rho} = \frac{\|\delta x^0\|}{\left(1 - \dfrac{1+\rho}{2} h_0\right)}$$

as stated above. Finally, upon inserting (2.46) into (2.47) and using $h_0^\delta \le h_0$, the result (2.48) is readily confirmed. □

Convergence analysis—`GBIT`. By a slight modification of Theorem 2.10, the Newton-`GBIT` iteration can also be shown to converge.

Theorem 2.11 *Let $\delta_k < \frac{1}{2}$ in (2.44) and replace the Kantorovich quantities h_k^δ in Theorem 2.10 by their upper bounds such that*

$$h_k^\delta = \frac{h_k}{1-\delta_k}\,.$$

Then we obtain the results:

I. Linear convergence mode. *Let δ_k in each inner iteration be controlled such that*

$$\vartheta(h_k, \delta_k) = \frac{\frac{1}{2}h_k^\delta + \delta_k(1 + h_k^\delta)}{1 - \delta_k} \leq \overline{\Theta}\,,$$

which assures that

$$\delta_k \leq \frac{\overline{\Theta}}{1 + \overline{\Theta}}\,. \tag{2.50}$$

Then this implies the inexact monotonicity test

$$\frac{\|\delta x^{k+1}\|}{\|\delta x^k\|} \leq \frac{1 - \delta_k}{1 - \delta_{k+1}}\overline{\Theta} \tag{2.51}$$

and the exact monotonicity test

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

II. Quadratic convergence mode. *Let the inner iteration be controlled according to* (2.47) *and*

$$h_0 < \frac{2(1 - \delta_0)^2}{1 + \rho}\,. \tag{2.52}$$

Then (2.48) *needs to be replaced by*

$$\rho > \frac{\delta_0(3 - 2\delta_0)}{1 - 2\delta_0}\,. \tag{2.53}$$

The exact Newton corrections behave like

$$\|\Delta x^{k+1}\| \leq \tfrac{1}{2}\frac{1 + \rho}{(1 - \delta_k)^2}\omega\|\Delta x^k\|^2$$

and the inexact Newton corrections like

$$\|\delta x^{k+1}\| \leq \tfrac{1}{2}\frac{1 + \rho}{1 - \delta_{k+1}}\omega\|\delta x^k\|^2\,.$$

Proof. The main difference to the previous theorem is that now we can only apply the triangle inequality

$$|\;\|\Delta x^k\| - \|\delta x^k - \Delta x^k\|\;| \leq \|\delta x^k\| \leq \|\delta x^k - \Delta x^k\| + \|\Delta x^k\|\,.$$

Assuming $\delta_k < 1$ in definition (2.44), we obtain

$$\frac{\|\Delta x^k\|}{1 + \delta_k} \leq \|\delta x^k\| \leq \frac{\|\Delta x^k\|}{1 - \delta_k}\,,$$

which motivates the majorant $\omega\|\delta x^k\| \leq h_k^\delta$ as stated in the theorem. Upon revisiting the proof of Theorem 2.10, the result (2.49) is seen to still hold, which is

$$\frac{\|\Delta x^{k+1}\|}{\|\delta x^k\|} \le \tfrac{1}{2} h_k^\delta + \delta_k (1 + h_k^\delta) \,. \tag{2.54}$$

From this, we obtain the modified estimate for the exact Newton corrections

$$\frac{\|\Delta x^{k+1}\|}{\|\Delta x^k\|} \le \frac{\tfrac{1}{2} h_k^\delta + \delta_k (1 + h_k^\delta)}{1 - \delta_k} = \frac{\tfrac{1}{2} h_k + \delta_k (1 - \delta_k + h_k)}{(1 - \delta_k)^2} = \vartheta(h_k, \delta_k) \,.$$

In a similar way, we obtain for the inexact Newton corrections

$$\frac{\|\delta x^{k+1}\|}{\|\delta x^k\|} \le \frac{\tfrac{1}{2} h_k^\delta + \delta_k (1 + h_k^\delta)}{1 - \delta_{k+1}} = \frac{1 - \delta_k}{1 - \delta_{k+1}} \vartheta(h_k, \delta_k) \,.$$

For the *linear* convergence mode, we adapt δ_k such that

$$\vartheta(h_k, \delta_k) \le \overline{\Theta} \,.$$

Asymptotically we thus assure that $\vartheta(0, \delta_k) \le \overline{\Theta}$, equivalent to (2.50).

For the *quadratic* convergence mode, we again require (2.47) (with h_k^δ in the present meaning, of course), i.e.

$$\delta_k \le \tfrac{1}{2} \rho \frac{h_k^\delta}{1 + h_k^\delta} \,.$$

With this choice we arrive at

$$\frac{\|\Delta x^{k+1}\|}{\|\Delta x^k\|} \le \tfrac{1}{2} \frac{1 + \rho}{1 - \delta_k} h_k^\delta = \tfrac{1}{2} \frac{1 + \rho}{(1 - \delta_k)^2} \omega \|\Delta x^k\|$$

for the *exact* Newton contraction, which requires (2.52) as a necessary condition. Upon combining (2.47) and (2.52), we obtain

$$\delta_0 \le \tfrac{1}{2} \rho \frac{h_0^\delta}{1 + h_0^\delta} < \frac{\rho(1 - \delta_0)}{1 + \rho + 2(1 - \delta_0)} \,.$$

Given ρ, this condition would lead to some uneasy quadratic root. Given δ_0, we merely have the linear inequality

$$\rho > \frac{\delta_0 (3 - 2\delta_0)}{1 - 2\delta_0} \,,$$

which is (2.53); it necessarily requires $\delta_0 < 1/2$ in agreement with the assumption $\delta_k < 1/2$ of the theorem.

The corresponding bound for the *inexact* Newton corrections is

$$\frac{\|\delta x^{k+1}\|}{\|\delta x^k\|} \le \tfrac{1}{2} \frac{1 + \rho}{1 - \delta_{k+1}} \omega \|\delta x^k\| \,,$$

which completes the proof. □

Convergence monitor. Assume that the quantity $\overline{\Theta} < 1$ in the linear convergence mode or the quadratic convergence mode have been specified; in view of (2.12), we may require that $\overline{\Theta} \leq 1/2$. The desirable convergence criterion would be

$$\Theta_k := \frac{\|\Delta x^{k+1}\|_2}{\|\Delta x^k\|_2} \leq \overline{\Theta}\,.$$

Since this criterion cannot be directly implemented, Θ_k needs to be substituted by a computationally available $\widetilde{\Theta}_k \approx \Theta_k$.

For CGNE with $\delta x_0^k = 0$, this leads to the inexact monotonicity test

$$\widetilde{\Theta}_k = \sqrt{\frac{1+\bar{\delta}_{k+1}^2}{1+\bar{\delta}_k^2}} \cdot \frac{\|\delta x^{k+1}\|_2}{\|\delta x^k\|_2} \leq \overline{\Theta}\,, \tag{2.55}$$

where the quantities $\bar{\delta}_k, \bar{\delta}_{k+1}$ are the computationally available estimates for the otherwise unavailable quantities δ_k, δ_{k+1} as given in (1.32).

For GBIT, the result (2.51) suggests the following inexact monotonicity test

$$\widetilde{\Theta}_k = \frac{1-\bar{\delta}_{k+1}}{1-\bar{\delta}_k} \cdot \frac{\|\delta x^{k+1}\|}{\|\delta x^k\|} \leq \overline{\Theta}\,. \tag{2.56}$$

As an alternative, we may also consider the weaker *necessary* condition

$$\widetilde{\Theta}_k = \frac{1-\bar{\delta}_{k+1}}{1+\bar{\delta}_k} \cdot \frac{\|\delta x^{k+1}\|}{\|\delta x^k\|} \leq \Theta_k \leq \overline{\Theta} \tag{2.57}$$

or the stronger *sufficient* condition

$$\Theta_k \leq \widetilde{\Theta}_k = \frac{1+\bar{\delta}_{k+1}}{1-\bar{\delta}_k} \cdot \frac{\|\delta x^{k+1}\|}{\|\delta x^k\|} \leq \overline{\Theta} \tag{2.58}$$

for use within the convergence monitor.

Preconditioning. In order to speed up the inner iteration, preconditioning from the left or/and from the right may be used. This means solving

$$\left(C_L F'(x^k) C_R\right) C_R^{-1}\left(\delta x^k - \Delta x^k\right) = C_L r^k\ .$$

In such a case, we will define

$$\delta_k = \frac{\|C_R^{-1}(\Delta x^k - \delta x^k)\|}{\|C_R^{-1}\delta x^k\|}\,.$$

Of course, in this case the preconditioned error norm is reduced by the inner iteration, whereas C_L only affects its rate of convergence. Consequently, any adaptive strategy should then, in principle, be based upon the contraction factors

$$\Theta_k = \frac{\|C_R^{-1}\Delta x^{k+1}\|}{\|C_R^{-1}\Delta x^k\|}$$

and its corresponding scaled estimate $\widetilde{\Theta}_k \approx \Theta_k$ as in (2.55) for CGNE or any choice between (2.56), (2.57), and (2.58) for GBIT.

Termination criterion. In the same spirit as above, we mimic the termination criterion (2.14) for the exact Newton iteration by requiring for `CGNE` the substitute condition

$$\frac{\sqrt{1+\bar{\delta}_k^2}}{1-\widetilde{\Theta}^2_{k-1}}\|\delta x^k\|_2 \le \mathrm{XTOL}$$

and for `GBIT` the sufficient condition

$$\frac{1+\bar{\delta}_k}{1-\widetilde{\Theta}^2_{k-1}}\|\delta x^k\| \le \mathrm{XTOL}\,,$$

each for the finally accepted iterate x^{k+1}, where XTOL is a user prescribed absolute *error tolerance* (to be replaced by some relative or some scaled error criterion).

Estimation of Kantorovich quantities. In order to deal successfully with the question of *how to match inner and outer iterations*, the above theory obviously requires the theoretical quantities $h_k^\delta = \omega\|\delta x^k\|$—which, however, are not directly available. In the spirit of the whole book we aim at replacing these quantities by *computational estimates* $[h_k^\delta]$. Recalling Section 2.1.1, we aim at estimating the a-priori estimates $[h_k] = 2\Theta^2_{k-1} \le h_k$ for $k \ge 1$.

For `CGNE` with initial correction $\delta x_0^k = 0$, we replace the relative errors δ_k by their estimates $\tilde{\delta}_k$ from Section 1.4.3 and thus arrive at the a-priori estimates

$$[h_k^\delta] = [h_k]/\sqrt{1+\bar{\delta}_k^2}\,, \quad [h_k] = 2\widetilde{\Theta}^2_{k-1} \le h_k\,, \quad k = 1, 2, \dots\,, \tag{2.59}$$

where $\widetilde{\Theta}_{k-1}$ from (2.55) is inserted.

For `GBIT`, we get the a-priori estimates

$$[h_k^\delta] = \frac{[h_k]}{(1-\bar{\delta}_k)}\,, \quad [h_k] = 2\widetilde{\Theta}^2_{k-1} \le h_k\,, \quad k = 1, 2, \dots\,, \tag{2.60}$$

where $\widetilde{\Theta}_{k-1}$ from (2.57) is inserted.

In both `CGNE` and `GBIT`, we may alternatively use the a-posteriori estimates

$$[h_{k-1}]_1 = 2\widetilde{\Theta}_{k-1}$$

and insert them either into (2.59) or into (2.60), respectively, to obtain $[h_{k-1}^\delta]_1$. From this, we may construct the a-priori estimates (for $k \ge 1$)

$$[h_k^\delta] = [h_{k-1}^\delta]_1\frac{\|\delta x^k\|}{\|\delta x^{k-1}\|}\,.$$

Note that in `CGNE` this formula inherits the saturation property.

For $k = 0$, we cannot but choose any ‘sufficiently small’ δ_0—as stated in the quadratic convergence mode to follow next.

Standard convergence mode. In this mode the inner iteration is terminated whenever

$$\delta_k \leq \bar{\delta} \tag{2.61}$$

for some default value $\bar{\delta} < 1$ to be chosen. In this case, *asymptotic linear convergence* is obtained.

For CGNE, Theorem 2.10 requires

$$\bar{\delta}/\sqrt{1+\bar{\delta}^2} < \overline{\Theta}\,,$$

which for $\overline{\Theta} = \frac{1}{2}$ leads to the restriction $\bar{\delta} < \sqrt{3}/3 \approx 0.577$. For GBIT, Theorem 2.11 requires

$$\bar{\delta}/(1-\bar{\delta}) < \overline{\Theta}\,,$$

which leads to $\bar{\delta} < 1/3$. In any case, we recommend to choose $\bar{\delta} \leq 1/4$ to assure at least two binary digits.

Quadratic convergence mode. In CGNE, we set $\delta_0 = \frac{1}{4}$ in (2.48) and obtain $\rho > 1$—thus assuring at least the first binary digit. In GBIT, we also set $\delta_0 = \frac{1}{4}$ and apply the inequality (2.53) thus arriving at $\rho > \frac{5}{4}$.

As for the adaptive termination of the inner iteration, we want to satisfy condition (2.47) for $k \geq 1$. Following our paradigm, we will replace the computationally unavailable quantity h_k^δ therein by its computational estimate $[h_k^\delta]$, which yields, for both CGNE and GBIT, the substitute condition

$$\bar{\delta}_k \leq \tfrac{1}{2}\rho \cdot \frac{[h_k^\delta]}{1+[h_k^\delta]}\,. \tag{2.62}$$

Whenever $\delta_k \leq \bar{\delta}_k$, the above monotone *increasing* right side as a function of $[h_k^\delta]$ and the relation $[h_k^\delta] \leq h_k^\delta$ imply that the theoretical condition (2.47) is actually *assured* with (2.62). Based on the a-priori estimates (2.59) or (2.60), respectively, we obtain a simple nonlinear scalar equation for an upper bound of δ_k.

Note that $\delta_k \to 0$ is enforced when $k \to \infty$, which means: *the closer the iterates come to the solution point, the more work needs to be done in the inner iteration to assure quadratic convergence of the outer iteration.*

Linear convergence mode. Once the approximated contraction factor $\widetilde{\Theta}_k$ is sufficiently below some prescribed threshold value $\overline{\Theta} \leq 1/2$, we may switch to the linear convergence mode described in either of the above two convergence theorems. As for the termination of the inner iteration, we recall the theoretical condition

$$\vartheta(h_k, \delta_k) \leq \overline{\Theta}\,.$$

Since the quantity ϑ is unavailable, we will replace it by the computationally available estimate

$$[\vartheta(h_k, \delta_k)] = \vartheta([h_k], \delta_k) \leq \vartheta(h_k, \delta_k)\,.$$

As this mode occurs only for $k > 0$, we can just insert the a-priori estimates (2.59) or (2.60), respectively. Since the above right hand side is a monotone *increasing* function of h_k and $[h_k] \leq h_k$, this estimate may be 'too small' and therefore lead to some δ_k, which is 'too large'. Fortunately, the difference between computational estimate and theoretical quantity can be ignored asymptotically. In any case, we require the monotonicity (2.55) for CGNE or (2.56), (2.57), or (2.58) for GBIT and run the inner iteration at each step k until either the actual value of δ_k obtained in the course of the inner iteration satisfies the condition above or divergence occurs with $\widetilde{\Theta}_k > 2\overline{\Theta}$.

In CGNE, we observe that in this mode *the closer the iterates come to the solution point, the less work is necessary within the inner iteration to assure linear convergence of the outer iteration.* In GBIT, this process continues only until the upper bound (2.50) for δ_k has been reached.

The here described error oriented local inexact Newton algorithms are self–contained and similar in spirit, but not identical with the local parts of the global inexact Newton codes GIANT-CGNE and GIANT-GBIT, which are worked out in detail in Section 3.3.4 below.

Bibliographical Note. A first affine covariant convergence analysis of a local inexact Newton method has been given by T.J. Ypma [203]. The first affine covariant inexact Newton code has been GIANT, developed by P. Deuflhard and U. Nowak [67, 160] in 1990. That code had also used a former version of GBIT for the inner iteration.

2.2 Residual Based Algorithms

In most algorithmic realizations of Newton's method iterative values of the *residual* norms are used for a check of convergence. An associated convergence analysis will start from *affine contravariant* Lipschitz conditions of the type (1.8) and lead to results in terms of residual norms only, which are tacitly assumed to be scaled. As explained in Section 1.2.2 above, such an analysis will not touch upon the question of local uniqueness of the solution.

2.2.1 Ordinary Newton method

Recall the notation of the ordinary Newton method

$$F'(x^k)\Delta x^k = -F(x^k)\,,\ x^{k+1} = x^k + \Delta x^k\,, \quad k = 0, 1, \ldots\,. \tag{2.63}$$

Convergence analysis. Analyzing the iterative residuals leads to an affine contravariant version of the well-known Newton-Mysovskikh theorem.

Theorem 2.12 *Let $F : D \to \mathbb{R}^n$ be a differentiable mapping with $D \subset \mathbb{R}^n$ open and convex. Let $F'(x)$ be invertible for all $x \in D$. Assume that the following affine contravariant Lipschitz condition holds:*

$$\left\|\big(F'(y) - F'(x)\big)(y-x)\right\| \le \omega \|F'(x)(y-x)\|^2 \text{ for } x, y \in D\,.$$

Define the open level set $\mathcal{L}_\omega = \left\{x \in D \mid \|F(x)\| < \frac{2}{\omega}\right\}$ and let $\overline{\mathcal{L}}_\omega \subset D$ be bounded. For a given initial guess x^0 of an unknown solution x^* let

$$h_0 := \omega \|F(x^0)\| < 2\,, \text{ i.e. } x^0 \in \mathcal{L}_\omega\,. \tag{2.64}$$

Then the ordinary Newton iterates $\{x^k\}$ defined by (2.63) remain in $\mathcal{L}_\omega$ and converge to some solution point $x^* \in \mathcal{L}_\omega$ with $F(x^*) = 0$. The iterative residuals $\{F(x^k)\}$ converge to zero at an estimated rate

$$\|F(x^{k+1})\| \le \tfrac{1}{2}\omega \|F(x^k)\|^2\,. \tag{2.65}$$

Proof. To show that $x^{k+1} \in D$ we apply the integral form of the mean value theorem and the above Lipschitz condition and obtain

$$\begin{aligned}
\|F(x^k + \lambda \Delta x^k)\| &= \|F(x^k) + \int\limits_{t=0}^{\lambda} F'(x^k + t\Delta x^k)\Delta x^k \, dt\| \\
&= \| \int\limits_{t=0}^{\lambda} \big(F'(x^k + t\Delta x^k) - F'(x^k)\big)\Delta x^k \\
&\quad +(1-\lambda)F(x^k)\, dt\| \\
&\le \int\limits_{t=0}^{\lambda} \|(F'(x^k + t\Delta x^k) - F'(x^k)\big)\Delta x^k\| \, dt \\
&\quad +(1-\lambda)\|F(x^k)\| \\
&\le \omega \int\limits_{t=0}^{\lambda} \|F'(x^k)\Delta x^k\|^2 t \, dt + (1-\lambda)\|F(x^k)\| \\
&= \big(1 - \lambda + \tfrac{1}{2}\omega\lambda^2 \|F(x^k)\|\big)\, \|F(x^k)\|
\end{aligned}$$

for each $\lambda \in [0,1]$ such that $x^k + t\Delta x^k \in \mathcal{L}_\omega$ for $t \in [0, \lambda]$. Now assume that $x^{k+1} \notin \mathcal{L}_\omega$. Then there exists a minimal $\overline{\lambda} \in]0,1]$ with $x^k + \overline{\lambda}\Delta x^k \in \partial\mathcal{L}_\omega$ and $\|F(x^k + \overline{\lambda}\Delta x^k)\| < (1 - \overline{\lambda} + \overline{\lambda}^2)\|F(x^k)\| < 2/\omega$, which is a contradiction. For $\lambda = 1$ we get relation (2.65). In terms of the residual oriented so-called Kantorovich quantities

$$h_k := \omega \|F(x^k)\| \tag{2.66}$$

we may obtain the quadratic recursion

$$h_{k+1} \le \tfrac{1}{2}h_k^2 = (\tfrac{1}{2}h_k)h_k\,. \tag{2.67}$$

With assumption (2.64), $h_0 < 2$, we obtain $h_1 < h_0 < 2$ for $k = 0$ and, by repeated induction over k, then

$$h_{k+1} < h_k < 2\,,\ k = 0, 1, \ldots \quad \Rightarrow \quad \lim_{k\to\infty} h_k = 0\,.$$

This can be also written in terms of the residuals as

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

In terms of the iterates we have

$$\{x^k\} \subset \mathcal{L}_\omega \subset D\,.$$

Since $\mathcal{L}_w$ is bounded, there exists an accumulation point x^* of $\{x^k\}$ with $F(x^*) = 0$, i.e. x^* is a solution point, but not necessarily unique in $\mathcal{L}_\omega$. □

This theorem also holds for underdetermined nonlinear systems—compare Exercise 4.10.

Convergence monitor. We now want to exploit Theorem 2.12 for actual computation. For this purpose, we introduce the contraction factors

$$\Theta_k := \frac{\|F(x^{k+1})\|}{\|F(x^k)\|}$$

and write (2.67) in the equivalent form

$$\Theta_k = \frac{h_{k+1}}{h_k} \leq \tfrac{1}{2} h_k\,. \tag{2.68}$$

For $k = 0$, assumption (2.64) assures *residual monotonicity*

$$\Theta_0 < 1\,. \tag{2.69}$$

Whenever $\Theta_0 \geq 1$, the assumption (2.64) is certainly violated, which means that the initial guess x^0 is not 'sufficiently close' to the solution point x^* in the sense of the above theorem. Suppose now that the test $\Theta_0 < 1$ has been passed. For the construction of a *quadratic convergence monitor* we introduce *computationally available estimates* $[h_k]$ for the unknown theoretical quantities h_k from (2.66). In view of (2.68) we may define the computational *a-posteriori* estimate

$$[h_k]_1 = 2\Theta_k \leq h_k$$

and, since $h_{k+1} = \Theta_k h_k$, also the *a-priori* estimate

$$[h_{k+1}] = \Theta_k [h_k]_1 = 2\Theta_k^2 \leq h_{k+1}\,.$$

Upon roughly identifying $[h_{k+1}]_1 \approx [h_{k+1}]$, we arrive at the approximate recursion ($k = 0, 1, \ldots$):

$$\Theta_{k+1} \approx \Theta_k^2 \leq \Theta_0 < 1\,.$$

Violation of this recursion at least in the mild sense

$$\Theta_{k+1} > \Theta_0$$

or the stricter sense

$$\Theta_{k+1} \geq 2\Theta_k^2$$

may be used to terminate the ordinary Newton iteration as 'not convergent'.

Termination criterion. This affine contravariant theory agrees with a termination criterion of the form

$$\|F(\hat{x})\| \leq \mathrm{FTOL}\,, \tag{2.70}$$

where FTOL is a user prescribed *residual error tolerance.*

Computational complexity. A short calculation shows that, for a given starting point x^0, the number q of iterations such that $\hat{x} = x^{q+1}$ meets the above termination requirement satisfies roughly

$$q \approx \mathrm{ld}\,\frac{\log(\mathrm{FTOL}\,/\|F(x^0)\|)}{\log \Theta_0}\,. \tag{2.71}$$

The proof is left as Exercise 2.1. In other words, with 'sufficiently good' initial guesses x^0 of the solution x^* at hand, the computational complexity of the nonlinear problem is comparable to the one of the linearized problem. Such problems are sometimes called *mildly nonlinear*.

2.2.2 Simplified Newton method

Recall the notation of the simplified Newton iteration

$$F'(x^0)\overline{\Delta x}^k = -F(x^k)\,,\ x^{k+1} = x^k + \overline{\Delta x}^k\,,\ k = 0, 1, \ldots\,. \tag{2.72}$$

Convergence analysis. Here we study convergence in terms of iterative residuals obtaining an affine contravariant variant of the Newton-Kantorovich theorem—without any uniqueness results, of course.

Theorem 2.13 *Let $F : D \to \mathbb{R}^n$ be $C^1(D)$ for $D \subset \mathbb{R}^n$ convex. Moreover, let $x^0 \in D$ denote a given starting point for the simplified Newton iteration* (2.72)*. Assume that the following affine contravariant Lipschitz condition holds:*

$$\left\|\left(F'(x) - F'(x^0)\right)v\right\| \leq \omega\|F'(x^0)(x - x^0)\| \cdot \|F'(x^0)v\| \tag{2.73}$$

for x, $x^0 \in D$, $v \in \mathbb{R}^n$ and $0 \leq \omega < \infty$. Define the level set

$$\mathcal{L}_\omega := \left\{x \in \mathbb{R}^n | \|F(x)\| \leq \frac{1}{2\omega}\right\}$$

and let $\overline{\mathcal{L}}_\omega \subseteq D$ be bounded. Assume that $x^0 \in \mathcal{L}_\omega$, which is

$$h_0 := \omega\|F(x^0)\| \leq \tfrac{1}{2}\,. \tag{2.74}$$

Then the iterates remain in $\mathcal{L}_\omega$ and converge to a solution point x^*. The iterative residual norms converge to zero at an estimated rate

$$\frac{\|F(x^{k+1})\|}{\|F(x^k)\|} \leq \tfrac{1}{2}(t_k + t_{k+1}) < 1 - \sqrt{1 - 2h_o}\,,$$

wherein the $\{t_k\}$ are defined by $t_0 = 0$ and

$$t_{k+1} = h_0 + \tfrac{1}{2}t_k^2\,,\ k = 0, 1, \ldots\,.$$

Proof. We apply the Lipschitz condition (2.73) to obtain

$$\begin{aligned}
\|F(x^{k+1})\| &= \left\|\int\limits_{t=0}^{1} \left(F'(x^k + t\overline{\Delta x}^k) - F'(x^0)\right)\overline{\Delta x}^k dt\right\| \\
&\leq \omega\|F'(x^0)\overline{\Delta x}^k\| \cdot \int\limits_{t=0}^{1} \|F'(x^0)(x^k - x^0 + t\overline{\Delta x}^k)\|dt
\end{aligned}$$

and, by triangle inequality:

$$\|F(x^{k+1})\| \leq \omega\|F(x^k)\|\left(\|F'(x^0)(x^k - x^0)\| + \tfrac{1}{2}\|F(x^k)\|\right). \tag{2.75}$$

We therefore introduce the *majorants*

$$\begin{aligned}
&\omega\|F'(x^0)(x^k - x^0)\| \leq t_k \\
&\omega\|F'(x^0)(x^{k+1} - x^k)\| = \omega\|F(x^k)\| \leq h_k
\end{aligned}$$

with initial values $t_0 = 0$, $h_0 \leq \frac{1}{2}$. Because of

$$\|F'(x^0)(x^{k+1} - x^0)\| \leq \|F'(x^0)(x^k - x^0)\| + \|F'(x^0)(x^{k+1} - x^k)\|$$

and the above relation (2.75), we obtain the same two majorant equations as in Section 2.1.2

$$t_{k+1} = t_k + h_k\,,\ h_k = h_{k-1}\left(t_{k-1} + \tfrac{1}{2}h_{k-1}\right)$$

and from these a single equation of the form

$$t_{k+1} - t_k = (t_k - t_{k-1})\left(t_{k-1} + \tfrac{1}{2}(t_k - t_{k-1})\right) = \tfrac{1}{2}(t_k^2 - t_{k-1}^2)\,.$$

Rearrangement of this equation permits the application of the *Ortega trick*

$$t_{k+1} - \tfrac{1}{2}t_k^2 = t_1 - \tfrac{1}{2}t_0^2 = h_0\,,$$

which once again may be interpreted as the simplified Newton iteration

$$t_{k+1} - t_k = -\frac{g(t_k)}{g'(t_0)} = g(t_k)$$

for the scalar equation

$$g(t) = h_0 - t + \tfrac{1}{2}t^2 = 0\,.$$

As can be seen from the above Figure 2.1, here also we obtain $g(t_{k+1}) < g(t_k)$, which is equivalent to $h_{k+1} < h_k$ and therefore

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

This assures that all simplified Newton iterates remain in $\mathcal{L}_\omega \subset D$. As for the convergence to some (not necessarily unique) solution point $x^* \in \mathcal{L}_\omega \subset D$, arguments similar to the ones used for Theorem 2.12 can be applied. As for the convergence rate, we go back to (2.75) and derive

$$\frac{\|F(x^{k+1})\|}{\|F(x^k)\|} \leq t_k + \tfrac{1}{2}h_k = \tfrac{1}{2}(t_k + t_{k+1}) < t^* = 1 - \sqrt{1 - 2h_0}\,,$$

which completes the proof. □

Convergence monitor. In order to exploit this theorem for actual implementation, we define the *residual contraction factors* ($k = 0, 1, \ldots$)

$$\Theta_k := \frac{\|F(x^{k+1})\|}{\|F(x^k)\|} \leq \tfrac{1}{2}(t_k + t_{k+1})\,.$$

For $k = 0$, the local convergence domain is characterized by

$$\Theta_0 \leq \tfrac{1}{2}h_0 \leq \tfrac{1}{4}\,, \tag{2.76}$$

which is clearly more restrictive than the comparable condition $\Theta_0 < 1$ for the ordinary Newton method—compare (2.69).

2.2.3 Broyden's 'bad' rank-1 updates

In this section, we deal with a quasi-Newton update already discussed by C.G. Broyden in his seminal paper [40] and classified there, on the basis of his numerical experiments, as being ‘bad’. This method can actually be derived in terms of affine contravariance. As stated before, only image space quantities like the residuals $F_k := F(x^k)$ are of interest in this frame. With $\delta F_{k+1} = F_{k+1} - F_k$, we rewrite the secant condition (1.17) here as

$$E_k(J)\delta F_{k+1} = F_{k+1} \tag{2.77}$$

in terms of the affine contravariant update change matrix

$$E_k(J) := I - J_k J^{-1} .$$

Any Jacobian rank-1 update satisfying

$$J_{k+1}^{-1} = J_k^{-1}\left(I - \frac{F_{k+1}v^T}{v^T\delta F_{k+1}}\right) , \; v \in R^n \; , \; v \neq 0$$

with v some vector in the image space of F will both satisfy the secant condition and reflect affine contravariance. As an example, the so-called ‘bad’ Broyden method is characterized by setting $v = \delta F_{k+1}$.

Convergence analysis. We start with an analysis of one quasi–Newton step of this kind.

Theorem 2.14 *Let*

$$J_{k+1}^{-1} = J_k^{-1}\left(I - \frac{F_{k+1}\delta F_{k+1}^T}{\|\delta F_{k+1}\|^2}\right) \tag{2.78}$$

denote the affine contravariant ‘bad’ Broyden rank-1 update and assume residual contraction

$$\Theta_k := \frac{\|F_{k+1}\|}{\|F_k\|} < 1 .$$

Then:

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

$$\begin{aligned} \|E_k(J_{k+1})\| &\leq \|E_k(J)\| \qquad \forall J \in S_k \\ \|E_k(J_{k+1})\| &\leq \frac{\Theta_k}{1-\Theta_k} . \end{aligned}$$

2. *The update matrix J_{k+1} is nonsingular whenever J_k is nonsingular and can be represented by*

$$J_{k+1} = \left(I - \frac{F_{k+1}\delta F_{k+1}^T}{\delta F_{k+1}^T F_k}\right) J_k \; .$$

3. *With* $\overline{\delta x}_{k+1} = -J_k^{-1} F_{k+1}$, *the next quasi-Newton correction is*

$$\delta x_{k+1} = -J_{k+1}^{-1} F_{k+1} = \left(1 - \frac{\delta F_{k+1}^T F_{k+1}}{\|\delta F_{k+1}\|^2}\right) \overline{\delta x}_{k+1} \,.$$

Proof. For the above rank-1 update we have

$$E_k(J_{k+1}) = \frac{F_{k+1} \delta F_{k+1}^T}{\|\delta F_{k+1}\|^2}$$

and therefore

$$\|E_k(J_{k+1})\| = \frac{\|E_k(J_{k+1})\delta F_{k+1}\|}{\|\delta F_{k+1}\|} = \frac{\|F_{k+1}\|}{\|\delta F_{k+1}\|} = \frac{\|E_k(J)\delta F_{k+1}\|}{\|\delta F_{k+1}\|} \leq \|E_k(J)\|$$

for all J satisfying the secant condition (2.77). Further, for $\Theta_k < 1$, we obtain

$$\|E_k(J_{k+1})\| = \frac{\|F_{k+1}\|}{\|\delta F_{k+1}\|} \leq \frac{\Theta_k}{1 - \Theta_k} \,,$$

which confirms the above statement 1. Statements 2 and 3 are direct consequences of the Sherman-Morrison formula.

□

The above Theorem 2.14 only deals with the situation within one iterative step. The iteration as a whole is studied next.

Theorem 2.15 *For* $F \in C^1(D)$, $F : D \subset R^n \to R^n$, *$D$ convex, let x^* denote a unique solution point of F with $F'(x^*)$ nonsingular. Assume that for some $\omega < \infty$ the affine contravariant Lipschitz condition*

$$\|(F'(x) - F'(x^*))(y - x)\| \leq \omega \, \|F'(x^*)(x - x^*)\| \, \|F'(x^*)(y - x)\| \qquad (2.79)$$

holds for $x, y \in D$. Consider the quasi-Newton iteration as defined in Theorem 2.14. *For some $\overline{\Theta}$ in the range* $0 < \overline{\Theta} < 1$ *assume that:*

1. *in terms of the affine contravariant* deterioration matrix

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

the initial approximate Jacobian satisfies

$$\overline{\eta}_0 := \|E_0\| < \overline{\Theta} \,,$$

2. *the initial guess x^0 satisfies*

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

Then the quasi-Newton iterates x^k converge to x^* in terms of errors as

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

or, in terms of residuals as

$$\|F_{k+1}\| \leq \overline{\Theta}\, \|F_k\|\,.$$

The convergence is superlinear with

$$\lim_{k\to\infty} \frac{\|F_{k+1}\|}{\|F_k\|} = 0\,. \tag{2.80}$$

As for the Jacobian rank-1 updates, the 'bounded deterioration property' holds in the form

$$\|E_k\| \leq \overline{\eta}_0 + \frac{t_0}{(1-t_0)(1-\overline{\Theta})} \leq \overline{\Theta}$$

together with the asymptotic property

$$\lim_{k\to\infty} \frac{\|E_k \delta F_{k+1}\|}{\|\delta F_{k+1}\|} = 0\,.$$

Proof. For ease of writing we characterize the Jacobian update approximation by

$$\eta_k := \frac{\|E_k \delta F_{k+1}\|}{\|\delta F_{k+1}\|}\,,\ \overline{\eta}_k := \|E_k\| \geq \eta_k\,.$$

For the convergence analysis we introduce

$$f_k := F'(x^*)(x^k - x^*) \text{ and } t_k := \omega\, \|f_k\|\,.$$

I. To begin with, we analyze the behavior of the iterative residuals:

$$\begin{aligned} F_{k+1} &= F_k + \int\limits_{s=0}^{1} F'(x^k + s\delta x_k)\delta x_k\, ds \\ &= \int\limits_{s=0}^{1} (F'(x^k + s\delta x_k) - F'(x^*))\delta x_k\, ds + (F'(x^*) - J_k)\delta x_k\,. \end{aligned}$$

Applying the Lipschitz condition (2.79) yields

$$\begin{aligned}
\|F_{k+1}\| &\leq \int\limits_{s=0}^{1} \|(F'(x^k + s\delta x_k) - F'(x^*))\delta x_k\| \, ds + \|(F'(x^*)J_k^{-1} - I)F_k\| \\
&\leq \int\limits_{s=0}^{1} \omega\|F'(x^*)(x^k + s\delta x_k - x^*)\| \, \|F'(x^*)\delta x_k\| \, ds + \|E_k F_k\| \\
&\leq \int\limits_{s=0}^{1} \omega\big(\|F'(x^*)(1-s)(x^k - x^*)\| \\
&\qquad + \|F'(x^*)s(x^{k+1} - x^*)\|\big) \, \|F'(x^*)\delta x_k\| \, ds + \overline{\eta}_k\|F_k\| \\
&= \tfrac{1}{2}(t_k + t_{k+1})\|F'(x^*)\delta x_k\| + \overline{\eta}_k\|F_k\| \, .
\end{aligned}$$

Defining $\bar{t}_k := \frac{1}{2}(t_k + t_{k+1})$, we get

$$\begin{aligned}
\|F_{k+1}\| &\leq \bar{t}_k\|(E_k - I)F_k\| + \overline{\eta}_k\|F_k\| \\
&\leq (\bar{t}_k(1 + \overline{\eta}_k) + \overline{\eta}_k)\|F_k\| \, .
\end{aligned} \tag{2.81}$$

As for the iterative errors f_k, we may derive the relation

$$\begin{aligned}
f_{k+1} &= f_k - F'(x^*)J_k^{-1}F_k = F'(x^*)(x^k - x^*) - F_k + E_k F_k \\
&= \int\limits_{s=0}^{1} \big(F'(x^*) - F'(x^* + s(x^k - x^*))\big) \, (x^k - x^*) \, ds + E_k F_k \, ,
\end{aligned}$$

from which we obtain the estimate

$$\begin{aligned}
\|f_{k+1}\| &\leq \int\limits_{s=0}^{1} s\omega \, \|F'(x^*)(x^k - x^*)\| \, \|F'(x^*)(x^k - x^*)\| \, ds + \overline{\eta}_k\|F_k\| \\
&\leq \frac{\omega}{2}\|f_k\|^2 + \overline{\eta}_k(\|f_k - F_k\| + \|f_k\|) \, .
\end{aligned}$$

By multiplication with ω and proceeding as above, this can be further reduced to yield

$$t_{k+1} \leq \tfrac{1}{2}t_k^2 + \overline{\eta}_k \left(\tfrac{1}{2}t_k^2 + t_k\right) = \left(\overline{\eta}_k + \frac{1 + \overline{\eta}_k}{2} t_k\right) t_k \, . \tag{2.82}$$

II. Next, we study the approximation properties of the Jacobian updates. Introducing the orthogonal projection

$$Q_k := \frac{\delta F_{k+1} \delta F_{k+1}^T}{\|\delta F_{k+1}\|^2}$$

onto the secant direction δF_{k+1}, the deterioration matrix may be written as

$$E_{k+1} = E_k Q_k^{\perp} + E_{k+1} Q_k \, , \tag{2.83}$$

yielding, as in the ‘good’ Broyden proof,

$$\overline{\eta}_{k+1} = \|E_{k+1}\| \le \|E_k Q_k^{\perp}\| + \|E_{k+1}Q_k\| \le \|E_k\| + \frac{\|E_{k+1}\delta F_{k+1}\|}{\|\delta F_{k+1}\|}\,.$$

Using the secant condition (2.77), we get for the numerator of the second right hand term:

$$\begin{aligned} E_{k+1}\delta F_{k+1} &= \delta F_{k+1} - F'(x^*)J_{k+1}^{-1}\delta F_{k+1} = \delta F_{k+1} - F'(x^*)\delta x_k \\ &= \int\limits_{s=0}^{1} (F'(x^k + s\delta x_k) - F'(x^*))\delta x_k\,. \end{aligned}$$

This can be estimated as above as follows

$$\begin{aligned} \|E_{k+1}\delta F_{k+1}\| &\le \bar{t}_k \|F'(x^*)\delta x_k\| \\ &= \bar{t}_k \|E_{k+1}\delta F_{k+1} - \delta F_{k+1}\| \\ &\le \bar{t}_k(\|E_{k+1}\delta F_{k+1}\| + \|\delta F_{k+1}\|) \end{aligned}$$

in order to get

$$\|E_{k+1}\delta F_{k+1}\| \le \frac{\bar{t}_k}{1-\bar{t}_k}\|\delta F_{k+1}\|\,. \tag{2.84}$$

Inserting this estimate into (2.83) yields the quite rough estimate

$$\overline{\eta}_{k+1} \le \overline{\eta}_k + \frac{\bar{t}_k}{1-\bar{t}_k}\,.$$

III. For the purpose of repeated induction assume that we have

$$\overline{\eta}_k \le \overline{\eta}_0 + \frac{\sum_{i=0}^{k-1}\overline{\Theta}^i t_0}{1-t_0} \le \overline{\eta}$$

with

$$\overline{\eta} := \overline{\eta}_0 + \frac{t_0}{(1-t_0)(1-\overline{\Theta})}$$

and

$$t_k \le \overline{\Theta}^k t_0\,.$$

Then by (2.82) and by the subsequent technical Lemma 2.16 below

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

and thus

$$\overline{\eta}_{k+1} \le \overline{\eta}_k + \frac{t_k}{1-t_0} \le \overline{\eta}_0 + \frac{\sum_{i=0}^{k-1}\overline{\Theta}^i t_0}{1-t_0} + \frac{\overline{\Theta}^{k+1}t_0}{1-t_0} \le \overline{\eta}_0 + \frac{\sum_{i=0}^{k}\overline{\Theta}^i t_0}{1-t_0} \le \overline{\eta}\,.$$

By induction we have the 'bounded deterioration property'

$$\overline{\eta}_k \leq \overline{\eta}$$

and the error contraction

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

for any k. Obviously, by (2.81) and the subsequent technical Lemma 2.16 we also have contraction of the residuals:

$$\|F_{k+1}\| \leq \overline{\Theta}\|F_k\|$$

IV. In order to show *superlinear* convergence, we use the orthogonal splitting provided by (2.83) in a more subtle manner. Since

$$Q_k E_k^T v = \delta F_{k+1} \frac{\langle \delta F_{k+1}, E_k^T v\rangle}{\|\delta F_{k+1}\|^2} = \delta F_{k+1} \frac{\langle E_k \delta F_{k+1}, v\rangle}{\|\delta F_{k+1}\|^2},$$

some short calculation supplies the equation

$$\begin{aligned}
\|E_{k+1}^T v\|^2 &= \|Q_k^\perp E_k^T v\|^2 + \|Q_k E_{k+1}^T v\|^2 \\
&= \|E_k^T v\|^2 - \|Q_k E_k^T v\|^2 + \|Q_k E_{k+1}^T v\|^2 \\
&= \|E_k^T v\|^2 - \frac{\langle E_k \delta F_{k+1}, v\rangle^2}{\|\delta F_{k+1}\|^2} + \frac{\langle E_{k+1} \delta F_{k+1}, v\rangle^2}{\|\delta F_{k+1}\|^2}.
\end{aligned}$$

Summing over the indices k, we arrive at

$$\sum_{k=0}^{l} \frac{\langle E_k \delta F_{k+1}, v\rangle^2}{\|\delta F_{k+1}\|^2 \|v\|^2} = \frac{\|E_0^T v\|^2}{\|v\|^2} - \frac{\|E_{l+1}^T v\|}{\|v\|^2} + \sum_{k=0}^{l} \frac{\langle E_{k+1} \delta F_{k+1}, v\rangle^2}{\|\delta F_{k+1}\|^2 \|v\|^2}.$$

Upon dropping the negative right hand term, letting $l \to \infty$, and using (2.84), we end up with the estimate

$$\sum_{k=0}^{l} \frac{\langle E_k \delta F_{k+1}, v\rangle^2}{\|\delta F_{k+1}\|^2 \|v\|^2} \leq \overline{\eta}_0^2 + \sum_{k=0}^{l} \left(\frac{t_k}{1-t_k}\right)^2 \leq \overline{\eta}_0^2 + \frac{t_0^2}{(1-t_0)^2(1-\overline{\Theta}^2)}.$$

Since the right hand side is bounded, we immediately conclude that

$$\lim_{k\to\infty} \frac{\langle E_k \delta F_{k+1}, v\rangle^2}{\|\delta F_{k+1}\|^2 \|v\|^2} = 0$$

for all $v \in R^n$. As a consequence, we must have

$$\lim_{k\to\infty} \eta_k = 0.$$

In order to prove the superlinear convergence statement (2.80), we may collect some estimates from above and proceed as

$$\begin{aligned} \|F'(x^*)J_k^{-1}F_{k+1}\| &= \|E_{k+1}\delta F_{k+1} - E_k\delta F_{k+1}\| \\ &\leq \bar{t}_k(1+\overline{\eta}_k)\|F_k\| + \eta_k\|\delta F_{k+1}\| \\ &\leq (\bar{t}_k(1+\overline{\eta}_k) + \eta_k(1+\overline{\Theta}))\|F_k\|\,. \end{aligned}$$

Finally, with

$$\begin{aligned} \|F_{k+1}\| - \|F'(x^*)J_k^{-1}F_{k+1}\| &\leq \|E_kF_{k+1}\| \leq \overline{\eta}_k\|F_{k+1}\| \\ \Rightarrow \qquad \|F_{k+1}\| &\leq \frac{\|F'(x^*)J_k^{-1}F_{k+1}\|}{1-\overline{\eta}_k}\,, \end{aligned}$$

we get

$$\|F_{k+1}\| \quad \leq \quad \frac{\bar{t}_k(1+\overline{\eta}_k) + \eta_k(1+\overline{\Theta})}{1-\overline{\eta\eta}_k}\|F_k\|\,.$$

Since $\bar{t}_k \to 0$ and $\eta_k \to 0$, superlinear convergence is easily verified. □

For ease of the above derivation, the following technical lemma has been postponed.

Lemma 2.16 *Assume* $0 < \Theta < 1$, $0 \leq \eta_0 < \Theta$ *and*

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

Then, with $\eta = \eta_0 + \dfrac{t}{(1-t)(1-\Theta)}$, *we have*

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

Proof. Under the given assumptions, a short calculation shows that $t < \frac{1}{7}$. Therefore we can proceed as

$$\begin{aligned} \Theta &\geq \eta_0 + \left(1 + \eta_0 + \tfrac{4}{3}(1-\Theta)^{-1}\right)t \\ &= \eta_0 + \frac{\frac{7}{6}t}{1-\Theta} + \left(1 + \eta_0 + \tfrac{1}{6}(1-\Theta)^{-1}\right)t \\ &\geq \eta_0 + \frac{t}{(1-t)(1-\Theta)} + \left(1 + \eta_0 + \frac{t}{(1-t)(1-\Theta)}\right)t \\ &= \eta + (1+\eta)t\,. \end{aligned}$$

□

Algorithmic realization. From representation (2.78) we again have a product form for the Jacobian update inverses. As a *condition number monitor* for the possible occurrence of ill-conditioning of the recursive Jacobian rank-1 updates, Lemma 2.8 may once more be applied, here to:

$$\mathrm{cond}_2(J_{k+1}) \leq \mathrm{cond}_2 \left(I - \frac{F_{k+1}\delta F_{k+1}^T}{\|\delta F_{k+1}\|^2} \right) \mathrm{cond}_2(J_k).$$

In the present context, we obtain for $\Theta_k < 1/2$:

$$\mathrm{cond}_2(J_{k+1}) \leq \frac{1}{1-2\Theta_k}\,\mathrm{cond}_2(J_k)\,.$$

As a consequence, a restriction such as

$$\Theta_k \leq \Theta_{\max} < \tfrac{1}{2}$$

with, say $\Theta_{\max} = 1/4$, will be necessary. With these preparations, we are now ready to present the 'bad Broyden' algorithm `QNRES` (the acronym stands for **RES**idual based **Q**uasi-**N**ewton algorithm).

Algorithm `QNRES`.

```
$F_0 := F(x^0)$                    evaluation and store
$\sigma_0 := \|F_0\|^2$                store
$J_0 \delta x_0 = -F_0$               linear system solve
$\kappa := 1$
For $k := 0, 1, \ldots, k_{\max}$:
    $x^{k+1} := x^k + \delta x^k$
    $F_{k+1} := F(x^{k+1})$
    $\delta F_{k+1} := F_{k+1} - F_k$
    $\sigma_{k+1} := \|F_{k+1}\|^2$
    If $\sigma_{k+1} \leq \mathrm{FTOL}^2$:
        solution found: $x^* = x^{k+1}$
    $\Theta_k := \sqrt{\sigma_{k+1}/\sigma_k}$
    If $\Theta_k \geq \Theta_{\max}$:
        stop: no convergence
    $w := \delta F_{k+1}$
    $\gamma_k := \|w\|^2$
    $\kappa := \kappa/(1 - 2\Theta_k)$
    If $\kappa \geq \kappa_{\max}$:
            **stop: ill-conditioned update**
    $v := (1 - \langle w, F_{k+1}\rangle/\gamma_k)F_{k+1}$
    For $j = k-1, \ldots, 0$:
            $\beta := \langle \delta F_{j+1}, v\rangle/\gamma_j$
            $v = v - \beta F_{j+1}$
    $J_0 \delta x_{k+1} = -v$
**stop: no convergence within $k_{\max}$ iterations**
```

The above algorithm merely requires to store the residuals $F_0, \ldots, F_{k+1}$, and the differences $\delta F_1, \ldots, \delta F_{k+1}$, which means an extra array storage of up to $2(k_{\max}+2)$ vectors of length n. Note that there is a probably machine-dependent tradeoff between computation and storage: the vectors δF_{j+1} can be either stored or recomputed. Moreover, careful considerations about *residual scaling* in the inner product $\langle \cdot, \cdot \rangle$ are recommended.

2.2.4 Inexact Newton-RES method

Recall inexact Newton methods with inner and outer iteration formally written as (dropping the inner iteration index i)

$$F'(x^k)\delta x^k = -F(x^k) + r^k\,,\ x^{k+1} = x^k + \delta x^k\,,\ k = 0, 1, \ldots\,. \tag{2.85}$$

In what follows, we will work out details for `GMRES` as inner iteration (see Section 1.4.1). For ease of presentation, we fix the initial values

$$\delta x_0^k = 0 \quad \text{and} \quad r_0^k = F(x^k)\,,$$

which, during the inner iteration ($i = 0, 1, \ldots$), implies in the generic case that

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

In what follows, we will denote the final value obtained from the inner iteration in each outer iteration step k by η_k, again dropping the inner iteration index i.

Convergence analysis. For the inexact Newton-`GMRES` iteration, we may state the following convergence theorem.

Theorem 2.17 *Let $F: D \to \mathbb{R}^n$, $F \in C^1(D)$, $D \subset \mathbb{R}^n$ convex. Let $x^0 \in D$ denote a given starting point for an inexact Newton iteration* (2.85)*. Assume the affine contravariant Lipschitz condition*

$$\left\|\bigl(F'(y)-F'(x)\bigr)(y-x)\right\| \le \omega\|F'(x)(y-x)\|^2$$
$$\text{for } 0 \le \omega < \infty\,, \quad \text{and } x,y \in D\,.$$

Let the level set $\mathcal{L}_0 := \bigl\{x \in \mathbb{R}^n |\ \|F(x)\| \le \|F(x^0)\|\bigr\} \subseteq D$ *be compact. For each well-defined iterate* $x^k \in D$ *define* $h_k := \omega\|F(x^k)\|$. *Then the outer residual norms can be bounded as*

$$\|F(x^{k+1})\| \le \bigl(\eta_k + \tfrac{1}{2}(1-\eta_k^2)h_k\bigr)\,\|F(x^k)\|\,. \tag{2.86}$$

The convergence rate can be estimated as follows:

I. Linear convergence mode. *Assume that the initial guess* x^o *gives rise to*

$$h_0 < 2\,.$$

Then some $\overline{\Theta}$ *in the range* $h_0/2 < \overline{\Theta} < 1$ *can be chosen. Let the inner* `GMRES` *iteration be controlled such that*

$$\eta_k \le \overline{\Theta} - \tfrac{1}{2}h_k\,. \tag{2.87}$$

Then the Newton-`GMRES` *iterates* $\{x^k\}$ *converge at least linearly to some solution point* $x^* \in \mathcal{L}_0$ *at an estimated rate*

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

II. Quadratic convergence mode. *If, for some* $\rho > 0$, *the initial guess* x^0 *guarantees that*

$$h_0 < 2/(1+\rho)$$

and the inner iteration is controlled such that

$$\frac{\eta_k}{1-\eta_k^2} \le \tfrac{1}{2}\rho h_k\,, \tag{2.88}$$

then the convergence is quadratic at an estimated rate

$$\|F(x^{k+1})\| \le \tfrac{1}{2}\omega(1+\rho)(1-\eta_k^2)\|F(x^k)\|^2\,. \tag{2.89}$$

Proof. Proceeding as in earlier proofs, we obtain

$$\begin{aligned}
\|F(x^{k+1})\| &= \|\int_0^1 \bigl(F'(x^k+t\delta x^k)-F'(x^k)\bigr)\delta x^k dt + r^k\| \\
&\le \int_0^1 \left\|\bigl(F'(x^k+t\delta x^k)-F'(x^k)\bigr)\delta x^k\right\| dt + \|r^k\| \\
&\le \tfrac{1}{2}\omega\|F(x^k)-r^k\|^2 + \|r^k\|\,.
\end{aligned}$$

By use of (1.20), this is seen to be just (2.86). Under the assumption (2.87) with $\overline{\Theta} < 1$ and $\eta_k < 1$ from `GMRES`, we obtain

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

and by repeated induction

$$\{x^k\} \subset \mathcal{L}_0 \subset D\,,$$

from which the convergence to $x^* \in \mathcal{L}_0$ is concluded. Quadratic convergence as in (2.89) is shown by mere insertion of (2.88) into (2.86). □

Convergence monitor. Throughout the inexact Newton iteration we will check for *residual monotonicity*

$$\Theta_k := \frac{\|F(x^{k+1})\|}{\|F(x^k)\|} \leq \overline{\Theta} < 1\,,\ k = 0, 1, \ldots\,,$$

introducing certain default parameters $\overline{\Theta}$ in accordance with the above Theorem 2.17. We will regard an iteration as *divergent*, whenever $\Theta_k \geq \overline{\Theta}$ holds.

Termination criterion. As in the exact Newton iteration, the finally accepted iterate $\hat{x}$ is required to satisfy

$$\|F(\hat{x})\| \leq \text{FTOL}$$

with FTOL a user prescribed *residual error tolerance.*

Standard convergence mode. If $\eta_k \leq \bar{\eta} < 1$ is prescribed by the user, then (2.86) implies that $\Theta_k \to \bar{\eta}$ and *asymptotic linear convergence* occurs—as already shown in the early pioneering paper [51].

Quadratic convergence mode. Assume that for $k = 0$ some value η_0 is prescribed; from numerical experiments, we know that this value should be sufficiently small—compare, e.g., Table 8.3 in Section 8.2 below. For $k \geq 0$, (2.89) suggests the *a-posteriori estimate*

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

and, since $h_{k+1} = \Theta_k h_k$, also the *a-priori estimate:*

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

For $k > 0$, shifting the index $k+1$ now back to k, we therefore require that

$$\frac{\eta_k}{1-\eta_k^2} \leq \tfrac{1}{2}\rho[h_k] \leq \tfrac{1}{2}\rho h_k\,, \tag{2.90}$$

which can be assured in the course of the iterative computation of δx^k and r^k. For the parameter ρ some value $\rho \approx 1$ seems to be appropriate. Note that asymptotically this choice leads to $\eta_k \to \rho[h_k] \to 0$.

Linear convergence mode. Once the local contraction factor Θ_k is sufficiently below some prescribed value $\overline{\Theta}$, we may switch to the linear convergence mode described in the above Theorem 2.17. Careful examination of the proof shows that

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

From this we may derive the *a-posteriori estimate*

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

and, since $h_{k+1} = \Theta_k h_k$, also the *a-priori estimate*

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

As a preparation of the next Newton step, we define

$$\overline{\eta}_{k+1} = \overline{\Theta} - \tfrac{1}{2}[h_{k+1}]$$

in terms of the above a-priori estimate. If this value is smaller than the value obtained from (2.90), then we continue the iteration in the quadratic convergence mode. Else, we realize the linear convergence mode in Newton step $k+1$ with some

$$\eta_{k+1} \leq \overline{\eta}_{k+1} \,.$$

Asymptotically, this strategy leads to $\eta_{k+1} \to \overline{\Theta}$.

Preconditioning. In order to speed up the inner iteration, preconditioning from the left or/and from the right may be used. This means solving

$$\left(C_L F'(x^k) C_R\right)\left(C_R^{-1}\delta x^k\right) = C_L\left(-F(x^k) + r^k\right)$$

instead of (2.85). In such a case, the norm of the *preconditioned residuals* $\bar{r}^k = C_L r^k$ is minimized in `GMRES`, whereas C_R only affects the rate of convergence via the Krylov subspace

$$\mathcal{K}_i(\bar{r}_0, \overline{A}) \text{ with } \overline{A} = C_L F'(x^k) C_R \,.$$

Consequently, the above strategy should be based on the contraction factors

$$\Theta_k = \frac{\|C_L F(x^{k+1})\|_2}{\|C_L F(x^k)\|_2}$$

for the outer iteration. Note, however, that C_L should *not* depend on the iterate x^k in this theoretical setting.

If strict residual minimization is aimed at, then only *right* preconditioning should be implemented (i.e., $C_L = I$).

The here described local Newton-`GMRES` algorithm is part of the global Newton code `GIANT-GMRES`, which will be described in Section 3.2.3 below.

Remark 2.2 If `GMRES` were replaced by some other *residual norm reducing* (but *not minimizing*) iterative linear solver, then a similar accuracy matching strategy can be worked out (left as Exercise 2.9).

Bibliographical Note. The concept of local inexact Newton methods—sometimes also called *truncated* Newton methods—seems to have first been published in 1982 by R.S. Dembo, S.C. Eisenstat, and T. Steihaug [51]; they presented an asymptotic analysis in terms of the residuals. In 1981, R.E. Bank and D.J. Rose [19] worked out details of an inexact Newton algorithm on the basis of residual control including certain algorithmic heuristics. In 1996, S.C. Eisenstat and H.F. Walker [91] suggested a further strategy to choose the η_k, which they call 'forcing terms'; their strategy is also based on convergence analysis results, but different from the one presented here.

2.3 Convex Optimization

In this section we consider the problem of minimizing a strictly convex functional $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^1$. Then $F(x) = f'(x)^T$ is a gradient mapping and $F'(x) = f''(x)$ is symmetric positive definite. We want to solve $F(x) = 0$, a system of n nonlinear equations, by local Newton methods. The convergence analysis will start from *affine conjugate* Lipschitz conditions of the type (1.9) and lead to results in terms of iterative functional values and energy norms of corrections or errors.

2.3.1 Ordinary Newton method

Recall the ordinary Newton method in the notation $(k = 0, 1, \ldots)$

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

Convergence analysis. We analyze its convergence behavior in terms of iterative values of the functional to be minimized and energy norms of the Newton corrections. Thus we arrive at an affine conjugate variant of the Newton-Mysovskikh theorem.

Theorem 2.18 *Let $f : D \to \mathbb{R}^1$ be a strictly convex C^2-functional to be minimized over some open and convex domain $D \subset \mathbb{R}^n$. Let $F(x) = f'(x)^T$ and $F'(x) = f''(x)$, which is symmetric and assumed to be strictly positive definite. Assume that the following affine conjugate Lipschitz condition holds:*

$$\big\|F'(z)^{-1/2}\big(F'(y)-F'(x)\big)(y-x)\big\| \leq \omega\|F'(x)^{1/2}(y-x)\|^2 \tag{2.91}$$

for collinear x, y, $z \in D$ with $0 \leq \omega < \infty$. For the initial guess x^0 assume that

$$h_0 = \omega\|F'(x^0)^{1/2}\Delta x^0\| < 2 \tag{2.92}$$

and that the level set $\mathcal{L}_0 := \{x \in D \mid f(x) \leq f(x^0)\}$ is compact. Then the ordinary Newton iterates remain in $\mathcal{L}_0$ and converge to the minimum point x^* at a rate estimated by

$$\|F'(x^{k+1})^{1/2}\Delta x^{k+1}\| \leq \tfrac{1}{2}\omega\|F'(x^k)^{1/2}\Delta x^k\|^2 \tag{2.93}$$

or, with $\epsilon_k := \|F'(x^k)^{1/2}\Delta x^k\|^2$ and $h_k := \omega\|F'(x^k)^{1/2}\Delta x^k\|$, by

$$\begin{array}{rclcl} -\frac{1}{6}h_k\epsilon_k & \leq & f(x^k) - f(x^{k+1}) - \frac{1}{2}\epsilon_k & \leq & \frac{1}{6}h_k\epsilon_k \\ \frac{1}{6}\epsilon_k & \leq & f(x^k) - f(x^{k+1}) & \leq & \frac{5}{6}\epsilon_k\,. \end{array} \tag{2.94}$$

The distance to the minimum can be bounded as

$$f(x^0) - f(x^*) \leq \frac{\frac{5}{6}\epsilon_0}{1 - h_0/2}\,.$$

Proof. With the Lipschitz condition (2.91) for $z = x^{k+1}$, $y = x^k + t\Delta x^k$, $x = x^k$, the result (2.93), which is equivalent to $h_{k+1} \leq h_k^2/2$, is proven just as before in Theorem 2.2. The fact that $x^{k+1} \in \mathcal{L}_0$ can be seen by applying the same technique as in the proof of Theorem 2.12 above. To derive (2.94), we verify that

$$\begin{array}{l} f(x^{k+1}) - f(x^k) + \frac{1}{2}\|F'(x^k)^{1/2}\Delta x^k\|^2 = \int\limits_{s=0}^{1} s \int\limits_{t=0}^{1} \langle \Delta x^k, w\rangle \, dtds\;, \\ \text{where } w = \big(F'(x^k + st\Delta x^k) - F'(x^k)\big)\Delta x^k \end{array} \tag{2.95}$$

with $\langle\cdot,\cdot\rangle$ the Euclidean inner product. The integrand term is estimated as

$$\begin{array}{rcl} \langle \Delta x^k\,,\; w\,\rangle & \leq & |\langle F'(x^k)^{1/2}\Delta x^k\,, F'(x^k)^{-1/2}w\,\rangle| \\ & \leq & \|F'(x^k)^{1/2}\Delta x^k\| \cdot \omega st\|F'(x^k)^{1/2}\Delta x^k\|^2 \end{array}$$

by the Cauchy-Schwarz inequality and (2.91) with $x = z = x^k$, $y = x^k + st\Delta x^k$. With $h_k < 2$ this is the left side of (2.94). Consequently, the iterates converge to x^*. Note that x^* is anyway unique in D under the assumptions made.

In order to obtain the right hand side of (2.94), we go up to (2.95), but this time apply Cauchy-Schwarz in the other direction, which yields:

$$0 \leq f(x^k) - f(x^{k+1}) \leq \left(\tfrac{1}{2} + \tfrac{1}{6}h_k\right)\|F'(x^k)^{1/2}\Delta x^k\|^2 < \tfrac{5}{6}\epsilon_k\,.$$

Summing over all $k = 0, 1, \ldots$ we get

$$0 \le \omega^2\bigl(f(x^0) - f(x^*)\bigr) \le \sum_{k=0}^{\infty} \bigl(\tfrac{1}{2}h_k^2 + \tfrac{1}{6}h_k^3\bigr) < \tfrac{5}{6}\sum_{k=0}^{\infty} h_k^2 \, .$$

By using

$$\tfrac{1}{2}h_{k+1} \le \bigl(\tfrac{1}{2}h_k\bigr)^2 \le \tfrac{1}{2}h_k < 1$$

the right hand upper bound can be further treated to obtain

$$\begin{aligned} (\tfrac{1}{2}h_0)^2 + (\tfrac{1}{2}h_1)^2 + \cdots \quad &\le \quad (\tfrac{1}{2}h_0)^2 + (\tfrac{1}{2}h_0)^4 + (\tfrac{1}{2}h_1)^4 + \cdots \\ &< \quad \tfrac{1}{4}h_0^2 \sum_{k=0}^{\infty} (\tfrac{1}{2}h_0)^k = \frac{\frac{1}{4}h_0^2}{1 - \frac{1}{2}h_0} \, , \end{aligned}$$

so that

$$\omega^2\bigl(f(x^0) - f(x^*)\bigr) < \frac{\frac{5}{6}h_0^2}{1 - \frac{1}{2}h_0} \, .$$

This is the last statement of the theorem. □

Convergence monitor. We now study the consequences of the above convergence theorem for actual implementation. Let ϵ_k, Θ_k be defined as

$$\epsilon_k = \|F'(x^k)^{1/2}\Delta x^k\|_2^2 = |\langle F(x^k), \Delta x^k\rangle| \, , \; \Theta_k = \left(\frac{\epsilon_{k+1}}{\epsilon_k}\right)^{1/2} \, .$$

Then the basic convergence result is

$$\Theta_k = \frac{h_{k+1}}{h_k} \le \tfrac{1}{2}h_k < 1$$

and

$$f(x^{k+1}) - f(x^k) < -\tfrac{1}{6}\epsilon_k \, .$$

For $k = 0$, we must have

$$\Theta_0 < 1$$

to assure that x^0 is within the local convergence domain. For $k > 0$, in a similar way as in the two cases before, we derive the approximate recursion ($k = 0, 1, \ldots$)

$$\Theta_{k+1} \approx \Theta_k^2 < \Theta_0 < 1 \, .$$

From this, we may terminate the iteration as 'divergent' whenever

$$f(x^{k+1}) - f(x^k) \ge -\tfrac{1}{6}\epsilon_k$$

or, since this criterion is prone to suffer from rounding errors, either

$$\Theta_k \ge \Theta_0 \quad (k > 0),$$

or

$$\Theta_{k+1} \ge \frac{\Theta_k^2}{\Theta_0} \, .$$

Termination criterion. We may terminate the iteration whenever either

$$\epsilon_k \le \mathrm{ETOL}^2$$

or, recalling that asymptotically

$$f(x^{k+1}) - f(x^k) \doteq -\tfrac{1}{2}\epsilon_k \,,$$

whenever

$$f(x^k) - f(x^{k+1}) \le \tfrac{1}{2}\,\mathrm{ETOL}^2$$

with ETOL a user prescribed *energy error tolerance.*

2.3.2 Simplified Newton method

Recall the notation of the simplified Newton iteration

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

Convergence analysis. We now want to study its functional minimization properties, when the Jacobian matrix is kept throughout the Newton iteration.

Theorem 2.19 *Let $f : D \to \mathbb{R}^1$ be a strictly convex C^2-functional to be minimized over some convex domain $D \subset \mathbb{R}^n$. Let $F(x) = f'(x)^T$ and $F'(x) = f''(x)$, which is then symmetric positive definite. Let $x^0 \in D$ be some given starting point for a simplified Newton iteration. Assume that the following affine conjugate Lipschitz condition holds:*

$$\|F'(x^0)^{-1/2}(F'(z) - F'(x^0))v\| \le \omega \|F'(x^0)^{1/2}(z - x^0)\| \cdot \|F'(x^0)^{1/2}v\|$$

for $z \in D$. Let

$$h_0 := \omega \|F'(x^0)^{1/2}\,\overline{\Delta x}^0\| \le \tfrac{1}{2}$$

and define $t^* = 1 - \sqrt{1 - 2h_0}$. Then, with $\epsilon_k := \|F'(x^0)^{1/2}\overline{\Delta x}^k\|^2$, the simplified Newton iteration converges to some x^* with

$$\omega \|x^* - x^0\| \le t^*\,.$$

The convergence rate can be estimated in terms of the functional by

$$-\tfrac{1}{6}\epsilon_k(t_{k+1} + 2t_k) \le f(x^k) - f(x^{k+1}) - \tfrac{1}{2}\epsilon_k \le \tfrac{1}{6}\epsilon_k(t_{k+1} + 2t_k) \tag{2.96}$$

or in terms of energy norms of the simplified Newton corrections by

$$\Theta_k = \left(\frac{\epsilon_{k+1}}{\epsilon_k}\right)^{1/2} \le \tfrac{1}{2}(t_{k+1} + t_k)\,,$$

wherein $\{t_k\}$ is defined from $t_0 = 0$ and

$$t_{k+1} = h_0 + \tfrac{1}{2}t_k^2 < t^*\,,\ k = 0, 1,\ \dots\,.$$

Proof. The proof is similar to the previous proofs of Theorem 2.5 and Theorem 2.13 and will therefore only be sketched here. With the definition for ϵ_k and the majorants

$$\omega \| F'(x^0)^{1/2}(x^k - x^0) \| \leq t_k \,, \; \omega \| F'(x^0)^{1/2}\, \overline{\Delta x}^k \| \leq h_k$$

we obtain for the functional decrease

$$\begin{aligned} & f(x^{k+1}) - f(x^k) + \tfrac{1}{2}\epsilon_k = \\ & \quad = \int\limits_{s=0}^{1} s \int\limits_{t=0}^{1} \left\langle \overline{\Delta x}^k ,\; \left(F'(x^k + ts\overline{\Delta x}^k) - F'(x^0) \right) \overline{\Delta x}^k \right\rangle dtds \\ & \quad \leq \omega\epsilon_k \int\limits_{s=0}^{1} s \int\limits_{t=0}^{1} \big((1-ts) \| F(x^0)^{1/2}(x^k - x^0) \| + \\ & \qquad + ts \| F'(x^0)^{1/2}(x^{k+1} - x^0) \| \big) \, dtds \\ & \quad \leq \tfrac{1}{6}\epsilon_k (t_{k+1} + 2t_k) \,. \end{aligned}$$

This is the basis for (2.96). The energy norm contraction factor arises as

$$\Theta_k = \left(\frac{\epsilon_{k+1}}{\epsilon_k} \right)^{1/2} \leq \tfrac{1}{2}(t_k + t_{k+1}) =: \frac{h_{k+1}}{h_k} \,.$$

With $t_0 = 0$, $t_{k+1} = t_k + h_k$ and the usual 'Ortega trick' the results above are essentially established. □

Convergence monitor. For actual computation, we also have

$$\Theta_0 \leq \tfrac{1}{2} h_0 \leq \tfrac{1}{4} \,.$$

Note that for the *simplified* Newton iteration, the asymptotic property $f(x^*) - f(x^k) \approx \frac{1}{2}\epsilon_k$ does *not* hold—compare (2.96). Mutatis mutandis, essentially just replacing norms by energy norms in the contraction factors Θ_k, the techniques already worked out in Section 2.1.2 carry over.

Termination criterion. This also can be directly copied from Section 2.1.2 with the proper replacement of norms by energy norms.

2.3.3 Inexact Newton-PCG method

We next study *inexact* Newton methods (dropping, as usual, the inner iteration index i)

$$F'(x^k)(\delta x^k - \Delta x_k) = r^k \,, \; x^{k+1} = x^k + \delta x^k \,, \qquad k = 0, 1, \ldots \,. \qquad (2.97)$$

In the context of (strictly) convex optimization the Jacobian matrices can be assumed to be symmetric positive definite, so that the outstanding candidate for an inner iteration will be the *preconditioned conjugate gradient* (`PCG`). Throughout this section we set $\delta x_0^k = 0$.

Convergence analysis. For the purpose of our analysis below, we recall the following *orthogonality condition*, which is equivalent to condition (1.21) independent of the selected preconditioner:

$$\langle \delta x^k,\ F'(x^k)(\delta x^k - \Delta x_k)\rangle = \langle \delta x^k, r^k\rangle = 0\,. \tag{2.98}$$

As before, Δx^k denotes the associated exact Newton correction. After these preparations, we are now ready to derive a Newton-Mysovskikh type theorem, which meets our above *affine conjugacy* requirements.

Theorem 2.20 *Let $f : D \to \mathbb{R}$ be a strictly convex C^2-functional to be minimized over some open and convex domain $D \subset \mathbb{R}^n$. Let $F'(x) := f''(x)$ be symmetric positive definite and let $\|\cdot\|$ denote the Euclidean vector norm. In the above introduced notation assume the existence of some $\omega < \infty$ such that the following affine conjugate Lipschitz condition holds for collinear x, y, $z \in D$:*

$$\left\|F'(z)^{-1/2}\big(F'(y) - F'(x)\big)v\right\| \le \omega\left\|F'(x)^{1/2}(y-x)\right\| \cdot \left\|F'(x)^{1/2}v\right\|.$$

Consider an inexact Newton-`PCG` *iteration* (2.97) *satisfying* (2.98) *and started with $\delta x_0^k = 0$. At any well-defined iterate x^k, define the exact Newton terms*

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

and, subject to inner iteration errors characterized by

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

the associated inexact Newton terms

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

For a given initial guess $x^0 \in D$ assume that the level set $\mathcal{L}_0 := \{x \in D \mid f(x) \le f(x^0)\}$ is closed and bounded. Then the following results hold:

I. Linear convergence mode. *Assume that x^0 satisfies*

$$h_0 < 2\overline{\Theta} < 2 \tag{2.99}$$

for some $\overline{\Theta} < 1$. Let $\delta_{k+1} \ge \delta_k$ throughout the inexact Newton iteration. Moreover, let the inner iteration be controlled such that

$$\vartheta(h_k^\delta, \delta_k) := \frac{h_k^\delta + \delta_k\left(h_k^\delta + \sqrt{4 + (h_k^\delta)^2}\,\right)}{2\sqrt{1+\delta_k^2}} \le \overline{\Theta}\,, \tag{2.100}$$

which assures that

$$\delta_k \le \overline{\Theta}/\sqrt{1-\overline{\Theta}^2}\,. \tag{2.101}$$

Then the iterates x^k remain in $\mathcal{L}_0$ and converge at least linearly to the minimum point $x^* \in \mathcal{L}_0$ such that

$$\|F'(x^{k+1})^{1/2}\Delta x^{k+1}\| \le \overline{\Theta}\,\|F'(x^k)^{1/2}\Delta x^k\| \tag{2.102}$$

and

$$\|F'(x^{k+1})^{1/2}\delta x^{k+1}\| \le \overline{\Theta}\,\|F'(x^k)^{1/2}\delta x^k\|\,.$$

II. Quadratic convergence mode. *Let for some $\rho > 0$ the initial iterate x^0 satisfy*

$$h_0^\delta < \frac{2}{1+\rho} \tag{2.103}$$

and the inner iteration be controlled such that

$$\delta_k \le \frac{\rho h_k^\delta}{h_k^\delta + \sqrt{4+(h_k^\delta)^2}}\,, \tag{2.104}$$

which requires that

$$\delta_0 < \frac{\rho}{1+\sqrt{1+(1+\rho)^2}}\,. \tag{2.105}$$

Then the inexact Newton iterates x^k remain in $\mathcal{L}_0$ and converge quadratically to the minimum point $x^* \in \mathcal{L}_0$ such that

$$\|F'(x^{k+1})^{1/2}\Delta x^{k+1}\| \le (1+\rho)\frac{\omega}{2}\|F'(x^k)^{1/2}\Delta x\|^2 \tag{2.106}$$

and

$$\|F'(x^{k+1})^{1/2}\delta x^{k+1}\| \le (1+\rho)\frac{\omega}{2}\|F'(x^k)^{1/2}\delta x\|^2\,. \tag{2.107}$$

III. Functional descent. *The convergence in terms of the functional can be estimated by*

$$-\tfrac{1}{6}h_k^\delta\epsilon_k^\delta \le f(x^k) - f(x^{k+1}) - \tfrac{1}{2}\epsilon_k^\delta \le \tfrac{1}{6}h_k^\delta\epsilon_k^\delta\,. \tag{2.108}$$

Proof. For the purpose of repeated induction, let $\mathcal{L}_k$ denote the level set defined in analogy to $\mathcal{L}_0$. First, in order to show that $x^{k+1} \in \mathcal{L}_k$, we start from the identity

$$\begin{aligned} &f(x^k + \lambda\delta x^k) - f(x^k) + \left(\lambda - \tfrac{1}{2}\lambda^2\right)\epsilon_k^\delta \\ &\quad = \int_{s=0}^{\lambda} s \int_{t=0}^{\lambda} \left\langle \delta x^k, \left(F'(x^k + st\delta x^k) - F'(x^k)\right)\delta x^k \right\rangle \,dt\,ds + \langle \delta x^k, r^k\rangle\,. \end{aligned}$$

The second right hand term vanishes due to (2.98). The energy product in the first term can be bounded as

$$\langle \delta x^k, \dots \rangle \le \|F'(x^k)^{1/2}\delta x^k\| \ \omega st \|F'(x^k)^{1/2}\delta x^k\|^2 = st h_k^\delta \epsilon_k^\delta \, .$$

For the purpose of repeated induction, let $h_k < 2$ and $\epsilon_k \neq 0$, which then implies that

$$f(x^k + \lambda \delta x^k) \le f(x^k) + \left(\tfrac{1}{3}\lambda^3 + \tfrac{1}{2}\lambda^2 - \lambda\right)\epsilon_k^\delta < f(x^k) \ \text{ for } \ \lambda \in]0,1] \, .$$

Therefore, the assumption $x^k + \delta x^k \notin \mathcal{L}_k$ would lead to a contradiction for some $\lambda \in]0,1]$.

For $\lambda = 1$, we get the left hand side of (2.108). Applying the Cauchy-Schwarz inequality in the other direction also yields the right hand side.

In order to monitor the behavior of the Kantorovich type quantities h_k, we estimate the local energy norms as

$$\begin{aligned} &\|F'(x^{k+1})^{1/2}\Delta x^{k+1}\| \\ &\quad \le \left\| F'(x^{k+1})^{-1/2} \left(\int\limits_{t=0}^{1} \left(F'(x^k + t\delta x^k) - F'(x^k)\right)\delta x^k dt + r^k \right) \right\| \\ &\quad \le \tfrac{1}{2}\omega \|F'(x^k)^{1/2}\delta x^k\|^2 + \|F'(x^{k+1})^{-1/2} r^k\| \, . \end{aligned}$$

With $z = \delta x^k - \Delta x^k$, the second right hand term can be estimated implicitly by

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

which leads to the explicit bound

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

Summarizing, we obtain the contraction factor bound

$$\Theta_k := \frac{\|F'(x^{k+1})^{1/2}\Delta x^{k+1}\|}{\|F'(x^k)^{1/2}\Delta x^k\|} \le \vartheta(h_k^\delta, \delta_k) \, . \tag{2.109}$$

Herein *linear* convergence shows up via (2.100) and (2.102). The result (2.101) is obtained with $h_k = 0$. Obviously, $h_k < 2\overline{\Theta}$ is necessary to obtain $\Theta_k \le \overline{\Theta}$ for some $\overline{\Theta} < 1$. As for the contraction of the inexact corrections, we apply $\delta_{k+1} \ge \delta_k$ and (1.26) to show that

$$\frac{\|F'(x^{k+1})^{1/2}\delta x^{k+1}\|}{\|F'(x^k)^{1/2}\delta x^k\|} = \sqrt{\frac{1+\delta_k^2}{1+\delta_{k+1}^2}}\,\Theta_k \le \Theta_k \le \overline{\Theta} \, .$$

Hence, we may complete the induction and conclude that the iterates x^k converge to x^*.

As for *quadratic* convergence, we impose condition (2.104) within (2.109) to obtain

$$\begin{aligned}\frac{\|F(x^{k+1})^{1/2}\Delta x^{k+1}\|}{\|F'(x^k)^{1/2}\Delta x^k\|} &\leq \frac{1}{2\sqrt{1+\delta_k^2}}\left(h_k^\delta + \delta_k(h_k^\delta + \sqrt{4+(h_k^\delta)^2}\right)\\ &\leq \tfrac{1}{2}(1+\rho)h_k^\delta\,,\end{aligned}$$

which, for $h_k^\delta \leq h_k \leq h_0$ assures the convergence relations (2.106) under the assumption (2.99). Upon inserting (2.103) into (2.104) we immediately verify (2.105). For the inexact corrections, we have equivalently

$$\begin{aligned}\frac{\|F(x^{k+1})^{1/2}\delta x^{k+1}\|}{\|F'(x^k)^{1/2}\delta x^k\|} &\leq \frac{1}{2\sqrt{1+\delta_{k+1}^2}}\left(h_k^\delta + \delta_k(h_k^\delta + \sqrt{4+(h_k^\delta)^2}\right)\\ &\leq \tfrac{1}{2}(1+\rho)h_k^\delta < 1\,,\end{aligned}$$

which then assures the convergence relations (2.107). This finally completes the proof. □

Convergence monitor. Assume now that we have a reasonable (and cheap) estimate of the relative energy norm errors δ_k available from the inner `PCG` iteration. A new iterate x^{k+1} might be accepted whenever either

$$f(x^{k+1}) - f(x^k) \leq -\tfrac{1}{6}\epsilon_k = -\tfrac{1}{6}(1+\delta_k^2)\epsilon_k^\delta\,.$$

or, as a slight generalization of the situation of Theorem 2.20, the *inexact monotonicity criterion*

$$\Theta_k := \left(\frac{\epsilon_{k+1}}{\epsilon_k}\right)^{1/2} = \left(\frac{(1+\delta_{k+1}^2)\epsilon_{k+1}^\delta}{(1+\delta_k^2)\epsilon_k^\delta}\right)^{1/2} \leq \overline{\Theta}_k < 1$$

holds. We will regard the outer iteration as *divergent*, if none of the above criteria is met.

Termination criteria. We will terminate the iteration whenever

$$\epsilon_k = (1+\delta_k^2)\epsilon_k^\delta \leq \mathrm{ETOL}^2 \quad \text{or} \quad f(x^k) - f(x^{k+1}) \leq \tfrac{1}{2}\,\mathrm{ETOL}^2\,. \tag{2.110}$$

Standard convergence mode. If we just impose the inner iteration termination criterion $\delta_k \leq \bar\delta$ for some fixed default value $\bar\delta$, we obtain *asymptotic linear convergence.* If we set $\overline{\Theta} = \frac{1}{2}$, then (2.101) induces $\bar\delta < \sqrt{3}/3$. As in the other two cases, we recommend $\bar\delta = 1/4$ to assure at least two binary digits.

Quadratic convergence mode. Assume that $[h_0] < 2/(1+\rho)$ for $\rho = 1$. Let δ_0 be given, say $\delta_0 = 1/4$ in agreement with (2.105). As for the adaptive termination of the inner iteration within the inexact local Newton method, we want to satisfy condition (2.104). Following our general paradigm, we will replace the unavailable upper bound therein by the computationally available condition in terms of computational estimates $[h_k]$ such that

$$\delta_k \leq \frac{\rho\,[h_k^\delta]}{[h_k^\delta] + \sqrt{4 + [h_k^\delta]^2}} \,. \tag{2.111}$$

Since the above right hand side is a monotone *increasing* function of $[h_k]$, the relation $[h_k] \leq h_k$ implies that the theoretical condition (2.104) is actually *assured* whenever (2.111) holds. Following our basic paradigm (compare Section 1.2), we apply (2.108) and define the computational *a-posteriori* estimates

$$[h_k^\delta]_2 = \frac{6}{\epsilon_k^\delta}|f(x^{k+1}) - f(x^k) + \tfrac{1}{2}\epsilon_k^\delta| \,, \quad [h_k]_2 = \sqrt{1+\delta_k^2}[h_k^\delta]_2 \,.$$

From this, shifting the index $k+1$ back to k, we may define the *a-priori* estimate

$$[h_k] = \Theta_{k-1}[h_{k-1}]_2, \tag{2.112}$$

which we insert into (2.111) to obtain a simple implicit scalar equation for δ_k.

Note that $\delta_k \to 0$ is forced when $k \to \infty$. In words: *the closer the iterates come to the solution point, the more work needs to be done in the inner iteration to assure quadratic convergence of the outer iteration.*

Linear convergence mode. Once the local contraction factor Θ_k is sufficiently below some prescribed value $\overline{\Theta}$, we may switch to the linear convergence mode described by the above Theorem 2.20. As for the termination of the inner iteration, we would like to assure condition (2.100), briefly recalled as

$$\vartheta(h_k^\delta, \delta_k) \leq \overline{\Theta} \,.$$

Since the above quantity ϑ is unavailable, we will replace it by the computationally available estimate

$$[\vartheta(h_k^\delta, \delta_k)] := \vartheta([h_k^\delta], \delta_k) \leq \vartheta(h_k^\delta, \delta_k) \,.$$

For $k > 0$, we may again insert the a-priori estimate (2.112) above. In any case, we will run the inner iteration until the actual δ_k satisfies either condition (2.100) for the linear convergence mode or condition (2.111) for the quadratic convergence mode. Whenever $\Theta_k \geq 1$ occurs, then we switch to some global variant of this local inexact Newton method—see Section 3.4.3.

Note that asymptotically

$$\delta_k \to \overline{\Theta}/\sqrt{1-\overline{\Theta}^2} \quad \text{as} \quad k \to \infty\,. \tag{2.113}$$

In other words: *the closer the iterates come to the solution point, the less work is necessary within the inner iteration to assure linear convergence of the outer iteration.*

The here described local inexact Newton algorithm for convex optimization is part of the global inexact Newton code `GIANT-PCG` worked out in detail in Section 3.4.3 below.

Bibliographical Note. The presentation in this chapter is a finite dimensional restriction of the affine conjugate convergence theory and the corresponding algorithmic concepts given by P. Deuflhard and M. Weiser [84] for nonlinear elliptic PDEs. Our here developed inexact Newton-`PCG` algorithm may be regarded as a competitor to nonlinear CG methods—both to the variant [93] due to R. Fletcher and C.M. Reeves and to the one due to E. Polak and R. Ribière [169, Section 2.3]. For the application of nonlinear CG to discrete partial differential equations see, e.g., the lecture notes [102] by R. Glowinski; from this perspective, our Newton-`PCG` method may be viewed as a nonlinear CG variant with Jacobian savings in a firm theoretical frame.

Exercises

Exercise 2.1 Derive the computational complexity bounds (2.71) in terms of number of iterations from Theorem 2.12.

Exercise 2.2 Let $M(x)$ denote a perturbed Jacobian matrix of the form $M(x^k) = F'(x^k) + \delta M(x^k)$. Derive a convergence theorem for a Newton-like method based on Theorem 2.10.

Exercise 2.3 As an illustration of the not affine covariant classical Newton-Mysovskikh theorem take $X = Y = \mathbb{R}^2$ and define

$$F(x) := \begin{pmatrix} x_1 - x_2 \\ (x_1 - 8)x_2 \end{pmatrix}.$$

Verify that here $h_F = \alpha_F \beta_F \gamma_F < 2$. The simple affine transformation

$$F \to G := \begin{pmatrix} 1 & 1 \\ 0 & \frac{1}{2} \end{pmatrix} F$$

induces the associated quantities α_G, β_G, γ_G, h_G. Once more, give best possible bounds and verify that now $h_G > 2$! Finally, prove that the affine invariant characterization from Theorem 2.2 yields $h_0 = \alpha\omega \ll 2$. Interpretation?

Hint: One obtains $h_F = 0.762$, $h_G = 2.159$, $h_0 = 0.127$.

Exercise 2.4 *Theorem of H.B. Keller.* Let $F : D \to \mathbb{R}^n$ be a continuously differentiable mapping with $D \subset \mathbb{R}^n$ convex. Suppose that $F'(x)$ is invertible for each $x \in D$ and satisfies the affine invariant Hölder continuity

$$\left\|F'(z)^{-1}\big(F'(y) - F'(x)\big)\right\| \leq \omega \|y - x\|^{\gamma} \,,$$

where $0 < \gamma \leq 1$.

a) Prove a variant of the affine covariant Newton-Mysovskikh theorem (Theorem 2.2).

b) Prove a variant of the affine covariant Newton-Kantorovich theorem (Theorem 2.1).

Exercise 2.5 *Theorem of L.B. Rall (improved by W.C. Rheinboldt).* Let $F : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$, D open convex. Assume that there exists a unique solution $x^* \in D$ and that $F'(x^*)$ is invertible. Let

$$\left\|F'(x^*)^{-1}\big(F'(y) - F'(x)\big)\right\| \leq \omega_* \|y - x\| \text{ for } \ x\,,\ y \in D$$

denote a special affine covariant Lipschitz condition. Let

$$S(x^*, \rho) := \{x \in X |\ \|x - x^*\| < \rho\} \subset D \,.$$

By introduction of the majorants

$$\omega_* \left\|x^k - x^*\right\| \leq t_k$$

prove that for any starting point $x^0 \in S(x^*, \rho)$ with $\rho := \dfrac{2}{3\omega_*}$, the ordinary Newton iteration remains in S and converges to x^*. Give a convergence rate estimate.

Exercise 2.6 For convex optimization there are three popular symmetric Jacobian rank-2 updates

- *Broyden-Fletcher-Goldfarb-Shanno* (BFGS):

$$J_{k+1} = J_k - \frac{F_k F_k^T}{\delta x_k^T J_k \delta x_k} + \frac{(F_{k+1} - F_k)(F_{k+1} - F_k)^T}{(F_{k+1} - F_k)^T \delta x_k} \,,$$

- *Davidon-Fletcher-Powell* (DFP):

$$\begin{aligned} J_{k+1} \quad &= \quad J_k + \frac{F_{k+1}(F_{k+1} - F_k)^T + (F_{k+1} - F_k)F_{k+1}^T}{(F_{k+1} - F_k)^T \delta x_k} - \\ &\quad - \frac{F_{k+1}^T \delta x_k}{((F_{k+1} - F_k)^T \delta x_k)^2} (F_{k+1} - F_k)(F_{k+1} - F_k)^T \,, \end{aligned}$$

- *Powell's symmetric Broyden* (PSB):

$$J_{k+1} = J_k + \frac{F_{k+1}\delta x_k^T + \delta x_k F_{k+1}^T}{\delta x_k^T \delta x_k} - \frac{F_{k+1}^T \delta x_k}{(\delta x_k^T \delta x_k)^2} \delta x_k \delta x_k^T \,.$$

a) Show that all updates satisfy the classical secant condition.

b) Which of these updates are defined in an affine conjugate way? For not affine conjugate updates: design an appropriate scaling so that at least scaling invariance is achieved.

c) Which of these updates can be interpreted as a least change secant update? Derive the associated error concept.

Exercise 2.7 *Rank-2 update formulas for convex optimization.* We consider several update formulas for convex optimization. Common basis for all these updates is the classical secant condition

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

a) Show that u and v in the general symmetric positive definite update formula

$$J = (I - uv^T)\, J_k (I - vu^T)$$

cannot be specified such that both the secant condition is satisfied and the update is of full rank 2.

b) Verify that this can be achieved by the comparable representation, the DFP update:

$$J_{k+1} = \left(I - \frac{\delta F_{k+1} \delta x_k^T}{(\delta F_{k+1}^T \delta x_k)} \right) J_k \left(I - \frac{\delta x_k \delta F_{k+1}^T}{(\delta F_{k+1}^T \delta x_k)} \right) + \frac{\delta F_{k+1} \delta F_{k+1}^T}{(\delta F_{k+1}^T \delta x_k)} \,.$$

c) Verify that this can be also achieved by the inverse representation, the BFGS update:

$$J_{k+1}^{-1} = \left(I - \frac{\delta x_k \delta F_{k+1}^T}{(\delta F_{k+1}^T \delta x_k)} \right) J_k^{-1} \left(I - \frac{\delta F_{k+1} \delta x_k^T}{(\delta F_{k+1}^T \delta x_k)} \right) + \frac{\delta x \delta x_k^T}{\delta F_{k+1}^T \delta x_k} \,.$$

Exercise 2.8 Recall the notation for quasi-Newton methods as given in Section 2.1.4. With the majorant definitions

$$\frac{\|\overline{\Delta x}_{k+1}\|}{\|\Delta x_k\|} \le \Theta_k < \tfrac{1}{2}\,, \ \ \|\Delta x_k\| \le e_k \,,$$

$$\left\| J_k^{-1} \left[F'(x^k) - J_k \right] \right\| \le \delta_k \,,$$

$$\left\| J_k^{-1} \left[F'(u) - F'(v) \right] \right\| \le \omega_k \, \|u - v\| \,,$$

verify the following set of recursions:

$$\begin{aligned}
\delta_{k+1} &= [\delta_k + \Theta_k + \omega_k e_k] \,/\, \big(1 - \Theta_k\big) \,, \\
e_{k+1} &= \frac{\Theta_k}{1-\Theta_k} e_k \,, \\
\omega_{k+1} &= \frac{\omega_k}{1-\Theta_k} \,, \\
\Theta_{k+1} &= \delta_{k+1} + \tfrac{1}{2}\omega_{k+1} e_{k+1} \,.
\end{aligned}$$

Under the additional assumption of 'bounded deterioration' in the form

$$\delta_k \le \delta$$

derive a Kantorovich-type local convergence theorem. Why is such a theorem unsatisfactory?

Exercise 2.9 Consider a residual based inexact Newton method, where the inner iteration is done by some *residual norm reducing*, but not *minimizing*, iterative solver—like the 'bad' Broyden algorithm BB for *linear* systems as described in [74]. Then the contraction results (2.86), which hold for the residual minimizer GMRES, must be replaced.

a) Show the alternative contraction result

$$\Theta_k \le \eta_k + \tfrac{1}{2}(1+\eta_k)^2 h_k \,.$$

b) For the Kantorovich quantities h_k, find cheap and reliable a-posteriori and a-priori computational estimates $[h_k] \le h_k$.

c) Design accuracy matching strategies (standard, linear, and quadratic convergence mode) similar to those worked out for GMRES in Section 2.2.4.

Exercise 2.10 Consider two Newton sequences $\{x^k\}$, $\{y^k\}$ starting at different initial guesses x^0, y^0 and continuing as

$$x^{k+1} = x^k + \Delta x^k \ , \quad y^{k+1} = y^k + \Delta y^k \ ,$$

where $\Delta x^k, \Delta y^k$ are the corresponding ordinary Newton corrections. Upon using the affine covariant Lipschitz condition

$$\|F'(u)^{-1}\left(F'(v) - F'(w)\right) u\| \le \omega \|v-w\| \|u\|$$

verify the nonlinear perturbation result

$$\|x^{k+1} - y^{k+1}\| \le \omega \left(\tfrac{1}{2}\|x^k - y^k\| + \|\Delta x^k\|\right) \|x^k - y^k\| \ .$$

Is the result invariant under $x \leftrightarrow y$?