Universität Karlsruhe
Rechenzentrum
Postfach 6980
76128 Karlsruhe
Germany

Interner Bericht Nr. 69/97

On Solvers for Nonlinear Large Systems

Abteilung "Numerikforschung für Supercomputer"
Rechenzentrum der Universität Karlsruhe
June 1997

Author: Rüdiger Weiss

On Solvers for Nonlinear Large Systems

Rüdiger Weiss*

Abstract

Nonlinear large systems of equations arising from the discretization of partial differential equations usually are solved by Newton's method in connection with iterative linear solvers. Alternatively, nonlinear cg approaches have been used. Nonlinear cg approaches like Fletcher-Reeves, Polak-Ribière or quasi Newton-type methods are commonly used for unconstrained minimization problems.

In this article different methods for the solution of nonlinear equations and minimization problems are classified and convergence estimates are derived. Several new methods for the solution of nonlinear systems are introduced. Numerical tests with a nonlinear system of partial differential equations confirm that the proposed methods converge faster than the classical Newton approach in connection with iterative linear solvers.

1 Background

Let us consider the two problems: to solve a system of nonlinear equations

$$f(x) = 0, \tag{1}$$

and to determine the argument x for the minimum of the function

$$\min_x F(x), \tag{2}$$

where $x \in I\!R^n$, $f : I\!R^n \to I\!R^n$ and $F : I\!R^n \to I\!R$. Let us assume that (1) and (2) have a unique solution. If F is differentiable, then the minimum (2) can be determined from the condition

$$f(x) = \nabla F(x) = 0.$$

Of course, there may be several local minima where the gradient vanishes. However, the minimization problem (2) is related to the problem (1) of finding a zero.

Equations of the type (2) arise in nonlinear and nonrestricted optimizations [10, 11], where the problems typically come from economics. The determination of eigenvalues by Reyleigh quotient minimization [5] is an example from

*Numerikforschung für Supercomputer, Rechenzentrum Universtät Karlsruhe, Postfach 6980, 76128 Karlsruhe, Germany, e-mail: weiss@rz.uni-karlsruhe.de

engineering. Our special interest is in the solution of partial differential equations (PDE). This means we deal with equation (1). A rough model for the 3-dimensional Navier-Stokes equations is

$$P(v) = \Delta v + R(v^T \nabla)v - h = 0, \tag{3}$$

where $P = (P_1, P_2, P_3)^T$, $v = (v_1, v_2, v_3)^T$ simulates a velocity and R simulates a Reynolds number. The number n of degrees of freedom is usually large for fine discretizations.

Usually, equation (1) is solved iteratively by linearizing techniques like the classical Newton's method or Broyden's method [2]. For Newton's method the iterate is determined from

$$x_m = x_{m-1} - [f'(x_{m-1})]^{-1} f(x_{m-1}) \tag{4}$$

with the Jacobian matrix

$$f' = \begin{pmatrix} \frac{\partial f^1}{\partial x^1} & \cdots & \frac{\partial f^1}{\partial x^n} \\ \vdots & & \vdots \\ \frac{\partial f^n}{\partial x^1} & \cdots & \frac{\partial f^n}{\partial x^n} \end{pmatrix},$$

where the superscript denotes the components of the vectors x and f. Instead of calculating x_m from (4) the linear systems

$$f'(x_{m-1})d_m = -f(x_{m-1}) \tag{5}$$

are solved in each iteration step in order to avoid a matrix inversion and

$$x_m = x_{m-1} + d_m. \tag{6}$$

If an iterative linear solver of CG-type (see for example [14]) is used for solving (5), then d_m is approximated by $d_{m,l}$ for $l = 0, \ldots$

$$d_{m,l} = \alpha_1 r_{m,l-1} + \sum_{i=2}^{\sigma_l} \alpha_i \left(d_{m,l-i+1} - d_{m,l-i}\right) + d_{m,l-1} \tag{7}$$

where $r_{m,l}$ is the residual of (5)

$$r_{m,l} = f'(x_{m-1})d_{m,l} + f(x_{m-1}). \tag{8}$$

Putting the outer Newton iteration together with the inner CG-type iteration we obtain in each inner iteration step aproximations

$$x_{m,l} = x_{m-1} + d_{m,l} \tag{9}$$

for the nonlinear equation (1) with

$$\begin{aligned} x_{m,l} &\overset{(9),(7)}{=} x_{m-1} + \alpha_1 r_{m,l-1} + \sum_{i=2}^{\sigma_l} \alpha_i \left(d_{m,l-i+1} - d_{m,l-i}\right) + d_{m,l-1} \\ &\overset{(9)}{=} \alpha_1 r_{m,l-1} + \sum_{i=2}^{\sigma_l} \alpha_i \left(x_{m,l-i+1} - x_{m,l-i}\right) + x_{m,l-1}. \end{aligned} \tag{10}$$

Substitution of the double indicees m, l by the simple index k gives

$$x_k \overset{(10)}{=} \alpha_1 r_{k-1} + \sum_{i=2}^{\sigma_k} \alpha_i \left(x_{k-i+1} - x_{k-i}\right) + x_{k-1}, \tag{11}$$

where

$$r_{k-1} = f'(x_{m-1})(x_{k-1} - x_{m-1}) + f(x_{m-1}) \tag{12}$$

for some $m < k$ determined by the outer Newton steps.

In each inner CG-type iteration step x_k in (11) converges to the solution x_m of (4), but not to the solution x of (1). Thus, an accurate solution of (4) by CG-type methods generally does not improve in the same way the solution of the original problem. Therefore, the linear solver is usually stopped after some digits of accuracy. On the other hand, even this moderate accuracy may require a large number of inner iterations. It is well known that for equation (3) a large number of iterations are needed of the iterative solvers GMRES, BCG, BICO or QMR (see [14]) if R increases and the corresponding discretization matrix becomes more and more non-symmetric. If the matrix is indefinite for some large value of R, then methods based on the normal equations like CGNE or CGNR (see [14]) have to be applied. These methods may require many iterations. Therefore, for large values of R often upwind differences are used for the first derivatives improving the convergence of linear solvers. However, in this case a large discretization error occurs.

Special methods for finding the minimum of (2) are the nonlinear cg versions of Fletcher-Reeves [7] and Polak-Ribière [12]. For both methods the new approximation is calculated from

$$x_k = \alpha q_k + x_{k-1} \tag{13}$$

so that α minimizes $F(\alpha q_k + x_{k-1})$. The search direction is

$$q_k = f(x_{k-1}) + \beta q_{k-1},$$

where $f(x) = \nabla F(x)$ is the gradient of F. The two nonlinear cg methods use a different β. The Fletcher-Reeves algorithm uses

$$\beta^{FR} = \frac{||f(x_{k-1})||^2}{||f(x_{k-2})||^2},$$

while the Polak-Ribière method takes

$$\beta^{PR} = \frac{f(x_{k-1})^T(f(x_{k-1}) - f(x_{k-2}))}{||f(x_{k-2})||^2}.$$

The nonlinear ORTHOMIN method by Chronopoulos [3] developed for the solution of (1) uses (13) where α is determined so that $||f(x_x)||$ is minimized and

$$q_k = f(x_{k-1}) + \sum_{i=1}^{\rho} \beta_i q_{k-i}.$$

The vectors q_k are orthogonal, i. e. $q_k^T q_{k-i} = 0$ for $i = 1, \ldots, \rho$.

Our aim is to develop solvers for (1) that converge more quickly and do not need essentially more storage or computational effort than the above mentioned classical techniques. In particular such approaches should be able to handle central difference approximations of (3) for large values of R without difficulties.

2 Nonlinear Approaches

All approximation methods that do not linearize equation (1) by Newton-like methods are considered to be nonlinear approaches. The first investigations of such methods are based on the work of and Fletcher and Powell [6]. Fletcher and Reeves [7] investigate nonlinear conjugate gradients (CG). Special aspects and modifications of this method are considered by Daniel [4] and by Gilbert and Nocedal [8]. The investigations of Chronopoulos [3] of a nonlinear ORTHOMIN method and of Axelsson and Chronopoulos [1] with nonlinear generalized CG are very promising.

In order to obtain a classification for very general methods we define in analogy to the linear case nonlinear orthogonalization methods.

Definition 2.1 An *orthogonalization method* is defined to be an iterative method satisfying the following relations. For $k \geq 1$

$$x_k \in \tilde{x}_k + \operatorname{span}(q_{k-\sigma_k,k}, \ldots, q_{k-1,k}), \tag{14}$$

where $\sigma_k \leq k$, $\tilde{x}_k \in \operatorname{span}(x_{k-\sigma_k}, \ldots, x_{k-1})$. The vectors $q_{k-i,k} \in \mathbb{R}^n$ are called *search directions.* The orthogonality condition

$$f(x_k)^T Z_k q_{k-i,k} = 0 \tag{15}$$

is satisfied for $i = 1, \ldots, \sigma_k$. Z_k are auxiliary, non-singular *orthogonalization matrices.*

The method is called

- *exact* if $\sigma_k = k$, i. e. k search directions $q_{k-i,k}$ are used for the calculation of the new iterate,
- *restarted* if $\sigma_k = (k-1) \bmod \sigma_{\text{res}} + 1$, i. e. a restart is made periodically after σ_{res} steps,
- *truncated* if $\sigma_k = \min(k, \sigma_{\max})$ with $\sigma_{\max}$ fixed, i. e. only $\sigma_{\max}$ $q_{k-i,k}$ are used for the calculation of the new iterate,
- *combined* if the truncated method is restarted,
i. e. $\sigma_k = \min((k-1) \bmod \sigma_{\text{res}} + 1, \sigma_{\max})$.

Definition 2.1 is precisely the same definition as for linear orthogonalization methods (see [14]) except that the linear expression $f(x_k) = Ax_k - b$ is substituted by a nonlinear function $f(x_k)$. It is clear that orthogonalization methods cover a large number of particular techniques. The difference is in the selection of the search directions, the orthogonalization matrices and the used strategy of restart and truncation.

From (14) follows that

$$x_k = \sum_{i=1}^{\sigma_k} \alpha_i q_{k-i,k} + \tilde{x}_k \tag{16}$$

with $\alpha_i \in I\!R$. If we choose $\tilde{x}_k = x_{k-1}$, $q_{k-1,k} = r_{k-1}$ as in (12), $q_{k-i,k} = x_{k-i+1} - x_{k-i}$ for $i = 2, \ldots, \sigma_k$, then we have precisely the form (11) for the new approximation. Thus, orthogonalization methods also describe the classical approach of Newton's method combined with iterative linear solvers. If the search directions are chosen as in (13), then orthogonalization methods describe nonlinear cg and generalized nonlinear cg methods. Moreover, orthogonalization methods cover methods that minimize the Euclidean norm of the function f. This approach is given in the next lemma.

Lemma 2.2 *If the orthogonalization matrix*

$$Z_k = f'(x_k) \tag{17}$$

is the Jacobian matrix, then

$$\|f(x_k)\| = \min_{\alpha_1,\ldots,\alpha_{\sigma_k}} \left\| f\left(\sum_{i=1}^{\sigma_k} \alpha_i q_{k-i,k} + \tilde{x}_k\right)\right\|. \tag{18}$$

Proof.

$$\begin{aligned}
\frac{\partial}{\partial \alpha_i}\|f(x_k)\|^2 &= 2f(x_k)^T \frac{\partial}{\partial \alpha_i} f(x_k) \\
&= 2f(x_k)^T f'(x_k) \frac{\partial}{\partial \alpha_i} x_k \\
&\overset{(16)}{=} 2f(x_k)^T f'(x_k) q_{k-i,k} \\
&\overset{(17)}{=} 2f(x_k)^T Z_k q_{k-i,k} \overset{(15)}{=} 0. \quad \square
\end{aligned}$$

From Lemma 2.2 follows directly that the exact solution of (1) is obtained at least in the step n if the search directions are linearly independent and if the orthogonalization method is exact. This result is only of theoretical interest because exact methods are in general not feasible in practice.

Lemma 2.2 is the key for the construction of the more general orthogonalization methods minimizing the Euclidean norm of the function f. The Jacobian matrix f' is needed according to Lemma 2.2 for the construction of such methods. We will first follow this approach and then introduce a more direct approach avoiding the calculation of the Jacobian with respect to x.

In a black-box environment it is possible that f is not available directly, but only the values of f and of the Jacobian f' can be calculated for some arguments x_k. In this case Lemma 2.2 provides a useful tool for constructing norm-minimizing methods. The coefficients α_i in (16) are determined from the nonlinear equations

$$f(x_k)^T f'(x_k) q_{k-i,k} = 0 \tag{19}$$

for $i = 1, \dots, \sigma_k$. Starting from the original nonlinear equation (1) the nonlinear equation (19) seems to be more complex. However, the dimension of the system (19) is σ_k and, therefore, relatively small compared to the dimension n of the original system.

We also remark that the solution of (19) needs not to be unique. In this case the choice of a particular solution may depend on the technique used for solving (19).

Equation (19) can be solved again by nonlinear orthogonalization methods (leading to a cascade of more more complex equations) or by Newton's method. If the latter method is applied, then the role of the two iterative methods of the classical approach (outer Newton iteration, inner CG-type iteration) is interchanged. The solution by Newton's method requires the calculation of the Jacobian matrix of (19) with respect to α_i.

Let $J_{i,j}$ be the entry i, j of the Jacobian $J(x_k)$ of (19) with respect to α_j, then

$$\begin{aligned}
J_{i,j} &= \frac{\partial}{\partial \alpha_j} \left(f(x_k)^T f'(x_k) q_{k-i,k} \right) \\
&= \left(\frac{\partial}{\partial \alpha_j} f(x_k) \right)^T f'(x_k) q_{k-i,k} + f(x_k)^T \left(\frac{\partial}{\partial \alpha_j} f'(x_k) \right) q_{k-i,k} \\
&= q_{k-j,k}^T f'(x_k)^T f'(x_k) q_{k-i,k} + f(x_k)^T \left[f''(x_k) q_{k-j,k} \right] q_{k-i,k},
\end{aligned}$$

where the components a, b of the matrix in the square bracket are

$$\left[f''(x_k) q_{k-j,k} \right]_{a,b} = \sum_{l=1}^{n} \frac{\partial^2 f^a}{\partial x^l \partial x^b} q_{k-j,k}^l . \tag{20}$$

Thus, the Jacobian matrix f' and the Hessian tensor f'' are needed if Newton's method is applied. Although the evaluation of these expressions by automatic differentiation methods (see [9]) may only need a number of operations comparable to the evalution of f, the main amount of work consists in determining these derivatives of f.

Let $\alpha = (\alpha_1, \dots, \alpha_{\sigma_k})^T$ and $q = (q_{k-1,k}, \dots, q_{k-\sigma_k,k})^T$, then in each iteration step the linear system of dimension σ_k

$$J(x_k)(\alpha_{new} - \alpha_{old}) = -q \left[f'(x_k) \right]^T f(x_k) \tag{21}$$

has to be solved, where x_k in (21) is evaluated according to (16) by using for α the old value α_{old}. Note that the matrix $J(x_k)$ is symmetric and positive definite and that the dimension of the system is small. Thus the direct Cholesky method or the iterative methods classical CG, ORTHOMIN or Fridman's method (see [14]) can be applied.

On the other hand, if the particular function f is known as for example in (3), then the evalution of the coefficients α_i in (16) can be done without calculating the Jacobian and Hessian tensor of f with respect to x_k. Let

$$\tilde{f}(\alpha) = f(x_k) = f\left(\sum_{i=1}^{\sigma_k} \alpha_i q_{k-i,k} + \tilde{x}_k\right). \tag{22}$$

Minimizing the Euclidean norm of f is equivalent to solving the nonlinear equations

$$\tilde{f}(\alpha)^T \frac{\partial}{\partial \alpha_i} \tilde{f}(\alpha) = 0 \tag{23}$$

for $i = 1, \ldots, \sigma_k$. For Newton's method the Jacobian matrix $\tilde{J}$ of (23) with respect to α is needed. This matrix has the entries

$$\tilde{J}_{i,j} = \left(\frac{\partial}{\partial \alpha_i} \tilde{f}(\alpha)\right)^T \frac{\partial}{\partial \alpha_j} \tilde{f}(\alpha) + \tilde{f}(\alpha)^T \frac{\partial^2}{\partial \alpha_i \partial \alpha_j} \tilde{f}(\alpha). \tag{24}$$

The new approximations for α are then computed by a formula similar to (21).

Let us apply this technique to the PDE (3) and let us assume for the sake of simple notations that v_k, y_{k-i} and $\tilde{v}_k$ are either discretized values corresponding to x_k, $q_{k-i,k}$ and $\tilde{x}_k$ or the continuous counterparts, i. e. functions with $x_k, q_{k-i,k}\tilde{x}_k : I\!R^3 \to I\!R^3$, Then we obtain the discretized or the continuous version of (16)

$$v_k = \sum_{i=1}^{\sigma_k} \alpha_i y_{k-i} + \tilde{v}_k \tag{25}$$

and similar to (22) for the operator P:

$$\begin{aligned} \tilde{P}(\alpha) &= \sum_{i=1}^{\sigma_k} \alpha_i \Delta y_{k-i} + \Delta \tilde{v}_k + R \sum_{i,j=1}^{\sigma_k} \alpha_i \alpha_j (y_{k-i}^T \nabla) y_{k-j} \\ &\quad + R \sum_{i}^{\sigma_k} \alpha_i \left[(y_{k-i}^T \nabla)\tilde{v}_k + (\tilde{v}_k^T \nabla) y_{k-i}\right] + R(\tilde{v}_k^T \nabla)\tilde{v}_k - h, \end{aligned} \tag{26}$$

$$\begin{aligned} \frac{\partial}{\partial \alpha_i} \tilde{P}(\alpha) &= \Delta y_{k-i} + R \sum_{j=1}^{\sigma_k} \alpha_j \left[(y_{k-i}^T \nabla) y_{k-j} + (y_{k-j}^T \nabla) y_{k-i}\right] \\ &\quad + R\left[(y_{k-i}^T \nabla)\tilde{v}_k + (\tilde{v}_k^T \nabla) y_{k-i}\right], \end{aligned} \tag{27}$$

$$\frac{\partial^2}{\partial \alpha_i \partial \alpha_j} \tilde{P}(\alpha) = R\left[(y_{k-i}^T \nabla) y_{k-j} + (y_{k-j}^T \nabla) y_{k-i}\right]. \tag{28}$$

Using equations (26) – (28) in a discretized version the above described technique of solving (23) can be applied. The continuous version also suggests to consider, instead of the numerical techniques, an algebraic version of orthogonalization methods. However, these considerations are far beyond the scope of this article.

If equations (26) – (28) are evaluated in the reverse order, then the second derivatives $\frac{\partial^2}{\partial\alpha_i\partial\alpha_j}\tilde{P}(\alpha)$ occur as terms in the formula (27) for $\frac{\partial}{\partial\alpha_i}\tilde{P}(\alpha)$:

$$\begin{aligned} \frac{\partial}{\partial\alpha_i}\tilde{P}(\alpha) &= \Delta y_{k-i} + \sum_{j=1}^{\sigma_k}\alpha_j\frac{\partial^2}{\partial\alpha_i\partial\alpha_j}\tilde{P}(\alpha) \\ &\quad + R\left[(y_{k-i}^T\nabla)\tilde{v}_k + (\tilde{v}_k^T\nabla)y_{k-i}\right], \end{aligned} \tag{29}$$

and these first derivatives $\frac{\partial}{\partial\alpha_i}\tilde{P}(\alpha)$ occur as terms in the formula (26) for $\tilde{P}(\alpha)$:

$$\begin{aligned} \tilde{P}(\alpha) &= \sum_{i=1}^{\sigma_k}\alpha_i\left[\frac{\partial}{\partial\alpha_i}\tilde{P}(\alpha) - \frac{1}{2}\sum_{j=1}^{\sigma_k}\alpha_j\frac{\partial^2}{\partial\alpha_i\partial\alpha_j}\tilde{P}(\alpha)\right] \\ &\quad + \Delta\tilde{v}_k + R(\tilde{v}_k^T\nabla)\tilde{v}_k - h. \end{aligned} \tag{30}$$

The derivatives can be considered as auxiliary values needed for the calculation of the function $\tilde{P}(\alpha)$. We conclude that the calculation of $\tilde{P}(\alpha)$ and the first and second derivatives is as expensive as if only $\tilde{P}(\alpha)$ would be computed.

3 Non-Minimizing Approach

We consider in this section orthogonalization methods with $Z_k = I$. In this case the nonlinear equations

$$f(x_k)^T q_{k-i,k} = 0, \tag{31}$$

for $1, \ldots, \sigma_k$ have to be solved. For Newton's method we get the the entries of the Jacobian J^{NM} from

$$J_{i,j}^{NM} = \left(\frac{\partial}{\partial\alpha_j}\tilde{f}(\alpha)\right)^T q_{k-i,k}. \tag{32}$$

One might be tempted to use such a method because the number of operations seems to be less than in the norm-minimizing case $Z_k = f'(x_k)$. However, a look at example (3) shows that this is not true. Although the second derivatives $\frac{\partial^2}{\partial\alpha_i\partial\alpha_j}\tilde{P}(\alpha)$ are not needed explicitely for (32) they are used implicitely for the calculation of $\frac{\partial}{\partial\alpha_i}\tilde{P}(\alpha)$; see (29). $J_{i,j}^{NM}$ in (32) seems to consist of less dot products than $\tilde{J}_{i,j}$ in (24). But $\tilde{J}_{i,j}$ is symmetric, so that only $\sigma_k^2 + \sigma_k$ dot products have to be calculated. On the contrary, $J_{i,j}^{NM}$ is non-symmetric, so that all σ_k^2 dot products have to be processed. The difference can be neglected.

As $J_{i,j}^{NM}$ is non-symmetric Gaussian elimination must be used if a direct solver is applied while for the symmetric $\tilde{J}_{i,j}$ the more economic Choleski method can be used. If the systems are solved iteratively usually non-symmtric systems require more iterations than symmetric systems. To sum it up, using $Z_k = I$ is not cheaper than using $Z_k = f'(x_k)$.

Moreover, no minimization result like in Lemma 2.2 or convergence estimate can be easily derived. Here is an essential difference to the linear orthogonalization methods (see [14]). In order to guarantee a monotonuous decrease of the norm of f an equivalent nonlinear residual smoothing (see [13]) has to be applied additionally. This leads to an orthogonalization method with $Z_k = f'(x_k)$ and $\sigma_k = 1$. Thus, we end up with a method descibed in the previous section.

For all these reasons we do not pursue this approach.

4 Search Directions

Throughout the rest of this paper we assume $Z_k = f'(x_k)$ and $\tilde{x}_k = x_{k-1}$. We look for one particular search direction that guarantees convergence. The next lemma is the basis for forthcoming convergence estimates.

Lemma 4.1 *Let q be any of the $q_{k-i,k}$ of a norm-minimizing orthogonalization method, then*

$$\|f(x_k)\|^2 \leq \|f(x_{k-1})\|^2 - \frac{\left(f(x_{k-1})^T f'(\xi) q\right)^2}{\|f'(\xi)q\|^2}, \tag{33}$$

where ξ is implicitly defined by $\xi = x_{k-1} + \tau q$ with $\tau = \theta\tau^*$, $0 < \theta < 1$ and $\tau^* = -\frac{f(x_{k-1})^T f'(\xi) q}{\|f'(\xi)q\|^2}$. τ^* minimizes $\|f\left(x_{k-1} + \tau^* q\right)\|$.

Proof.

$$\begin{aligned}
\|f(x_k)\|^2 &\overset{(18)}{=} \min_{\alpha_1,\ldots,\alpha_{\sigma_k}} \left\| f\left(\sum_{i=1}^{\sigma_k} \alpha_i q_{k-i,k} + x_{k-1}\right)\right\|^2 \\
&\leq \min_{\tau^*} \|f\left(\tau^* q + x_{k-1}\right)\|^2 \\
&= \min_{\tau^*} \|f(x_{k-1}) + \tau^* f'(\xi) q\|^2
\end{aligned}$$

by the mean value theorem with $\xi = x_{k-1} + \tau q$ and $\tau = \theta\tau^*$ for some $0 < \theta < 1$. The last expression is minimized by

$$\tau^* = -\frac{f(x_{k-1})^T f'(\xi) q}{\|f'(\xi)q\|^2}.$$

Inserting this τ^* into the last inequality we finally obtain the estimate (33). $\square$

Lemma 4.1 is the key for the choice of one particular q guaranteeing a convergence. From Lemma 4.1 follows directly that it is optimal if one of the search directions is the solution of the linear system

$$f'(\xi)q = f(x_{k-1}). \tag{34}$$

In this case we obtain from (33)

$$\|f(x_k)\|^2 \leq \|f(x_{k-1})\|^2 - \frac{\left(f(x_{k-1})^T f(x_{k-1})\right)^2}{\|f(x_{k-1})\|^2} = 0.$$

However, ξ cannot be determined easily from the implicit definition and the solution of (34) is the same problem as the solution of the Newton correction in the classical approach. If $f'(\xi)$ is approximated by $f'(x_{k-1})$ in (34) and $[f'(x_{k-1})]^{-1}$ is approximated by a truncated polynomial expansion (like for example a Neumann series), then we obtain search directions of the form

$$q = (\rho_1 f'(x_{k-1}) + \rho_0 I)\, f(x_{k-1}). \tag{35}$$

In linear CG such a search direction accelerates in general only the convergence if the coefficients ρ_i are chosen from a minimization principle. Thus, it is close at hand to consider the two directions in (35) separately and to construct norm-minimizing nonlinear solvers with two search directions so that the coefficients are determined by the norm minimizing process. This leads to the two search directions

$$q = f'(x_{k-1})f(x_{k-1}), \tag{36}$$
$$q = f(x_{k-1}). \tag{37}$$

For the linear CG-type methods usually the residuals are used as search directions. This corresponds for the nonlinear equation (1) to the choice (37). If the Jacobian would be the unit matrix, then this choice would be optimal according to (34). In linear CG this choice is motivated by the fact that the residual is the gradient of the quadratic funtional

$$E(z) = f(z)^T \left[f'(z)\right]^{-1} f(z).$$

The gradient is orthogonal to the isolines $E(z) = E(x_{k-1})$ and, therefore, the search direction is in the direction of the steepest descent. For a nonlinear f the gradient of the quadratic functional E involves the Hessian tensor and is very complex.

If we consider that our approach minimizes the quadratic funtional

$$F(z) = f(z)^T f(z),$$

then it is close at hand to choose the gradient of F as search direction. This gradient evaluated at the approximation x_{k-1} gives the search direction

$$q = \left[f'(x_{k-1})\right]^T f(x_{k-1}). \tag{38}$$

In the linear case this choice leads to the solvers CGNE and CGNR.

5 Function Values and Errors

Before the different choices of q are investigated auxiliary estimates are given needed for forthcoming theorems. Of particular interest is the connection between function values and errors.

Lemma 5.1 *For the relative errors and function values of any function f with $f(x) = 0$*

$$\frac{\|x_k - x\|}{\|x_0 - x\|} \leq \|f'(\xi_0)\| \left\| [f'(\xi_k)]^{-1} \right\| \frac{\|f(x_k)\|}{\|f(x_0)\|} \tag{39}$$

is satisfied for $\xi_i = x + \theta_i(x_i - x)$ with $0 < \theta_i < 1$ and $i = 1, k$.

Proof. By the mean value theorem we obtain for $\xi_i = x + \theta_i(x_i - x)$ with $0 < \theta_i < 1$

$$f(x_i) \quad = \quad f(x + x_i - x) = f(x) + f'(\xi_i)(x_i - x) \overset{f(x)=0}{=} f'(\xi_i)(x_i - x).$$

Evaluating this expression for $i = 0$ and taking the Euclidean norm gives

$$\|f(x_0)\| \quad \leq \quad \|f'(\xi_0)\| \ \|(x_0 - x)\|.$$

Similarly we obtain for $i = k$

$$\|(x_k - x)\| \quad \leq \quad \left\| [f'(\xi_k)]^{-1} \right\| \ \|f(x_k)\|.$$

The statement is proved multiplying the second inequality by the first. □

This lemma is equivalent to the linear case where errors and function values, in this case residuals, are connected by the condition number of the matrix (see [14]).

6 Convergence Estimates

In this section the different above mentioned search directions are theoretically investigated. In particular estimates for the three search directions in (37), (36) and (38) are derived. Thus, we consider the directions $q = f(x_{k-1})$, $q = f'(x_{k-1})f(x_{k-1})$ and $q = [f'(x_{k-1})]^T f(x_{k-1})$.

The next lemmata are needed in forthcoming proofs.

Lemma 6.1 *Let $A \in \mathbb{R}^{n\times}$ be an arbitrary square matrix. Then A can be split into the symmetric part M and the skew-symmetric part S with*

$$A = M + S, \ M^T = M, \ S^T = -S.$$

For any vector $z \in \mathbb{R}^n$ the following equations are valid:

$$z^T S z = 0, \tag{40}$$
$$z^T A^2 z = z^T M^2 z + z^T S^2 z. \tag{41}$$

If $\mu > 0$ exists such that for any $z \in \mathbb{R}^n$

$$z^T A z \geq \mu \|z\|^2, \tag{42}$$

then

$$\left\|A^{-1}\right\| \leq \frac{1}{\mu}, \tag{43}$$

$$z^T M z \geq \mu \|z\|^2, \tag{44}$$

$$z^T M^2 z \geq \mu^2 \|z\|^2, \tag{45}$$

and for any $w \in \mathbb{R}^n$

$$\frac{\|Aw\|}{\|w\|} \geq \mu. \tag{46}$$

Proof. Trivial. □

The next lemma is an estimate for the norm of the function values and it is more specific than Lemma 4.1.

Lemma 6.2 *$\mu > 0$ and $\omega > 0$ exist such that for any $z \in \mathbb{R}^n$ and for all vectors ξ on the straight line $\xi = x_{k-1} + \tau q$ with $\tau = \theta\tau^*$, where $0 < \theta < 1$, and τ^* minimizes $\|f(x_{k-1} + \tau^* q)\|$,*

$$z^T f'(\xi) z \geq \mu \|z\|^2, \tag{47}$$

$$\|\left[f''(\xi)z\right] z\| \leq \omega \|z\|^2. \tag{48}$$

Then for all norm-minimizing orthogonalization methods, where q is any of the $q_{k-i,k}$,

$$\|f(x_k)\|^2 \leq \|f(x_{k-1})\|^2 - \frac{1}{\left(1 + \frac{\|f(x_{k-1})\|\omega}{\mu^2}\right)^2} \frac{\left(f(x_{k-1})^T f'(x_{k-1})q\right)^2}{\|f'(\xi)q\|^2} \tag{49}$$

is valid.

Proof. For $\xi = x_{k-1} + \theta\tau^* q \overset{\text{Lemma 4.1}}{=} x_{k-1} - \theta \frac{f(x_{k-1})^T f'(\xi)q}{\|f'(\xi)q\|^2} q$ with $0 < \theta < 1$ we obtain by the mean value theorem

$$f'(\xi) = f'(x_{k-1}) - \theta \frac{f(x_{k-1})^T f'(\xi)q}{\|f'(\xi)q\|^2} \left[f''(\tilde{\xi})q\right]$$

with $\tilde{\xi} = x_{k-1} - \tilde{\theta}\theta \frac{f(x_{k-1})^T f'(\xi)q}{\|f'(\xi)q\|^2} q$ and $0 < \tilde{\theta} < 1$. Thus,

$$f(x_{k-1})^T f'(\xi)q = f(x_{k-1})^T f'(x_{k-1})q - \theta \frac{f(x_{k-1})^T f'(\xi)q}{\|f'(\xi)q\|^2} f(x_{k-1})^T \left[f''(\tilde{\xi})q\right] q.$$

Solving this equation for $f(x_{k-1})^T f'(\xi)q$ with the abbreviations $h = \left[f''(\tilde{\xi})q\right] q$ and $f = f(x_{k-1})$ gives

$$f^T f'(\xi)q = \frac{1}{1 + \theta \frac{f^T h}{\|f'(\xi)q\|^2}} f^T f'(x_{k-1})q. \tag{50}$$

From Lemma 4.1 follows

$$\begin{aligned}
\|f(x_k)\|^2 \quad & \leq \quad \|f\|^2 - \frac{\left(f^T f'(\xi)q\right)^2}{\|f'(\xi)q\|^2} \\
& \overset{(50)}{=} \quad \|f\|^2 - \frac{1}{\left(1 + \theta \frac{f^T h}{\|f'(\xi)q\|^2}\right)^2} \frac{\left(f^T f'(x_{k-1})q\right)^2}{\|f'(\xi)q\|^2} \\
& \overset{0<\theta<1}{\leq} \quad \|f\|^2 - \frac{1}{\left(1 + \frac{\|f\|\ \|h\|}{\|f'(\xi)q\|^2}\right)^2} \frac{\left(f^T f'(x_{k-1})q\right)^2}{\|f'(\xi)q\|^2} \\
& \overset{(48)}{\leq} \quad \|f\|^2 - \frac{1}{\left(1 + \frac{\|f\|\ \omega\ \|q\|^2}{\|f'(\xi)q\|^2}\right)^2} \frac{\left(f^T f'(x_{k-1})q\right)^2}{\|f'(\xi)q\|^2} \\
& \overset{(46)}{\leq} \quad \|f\|^2 - \frac{1}{\left(1 + \frac{\|f\|\ \omega}{\mu^2}\right)^2} \frac{\left(f^T f'(x_{k-1})q\right)^2}{\|f'(\xi)q\|^2}.
\end{aligned}$$

The last inequality is valid because f' is positive real by (47) so that inequality (46) can be applied. $\square$

6.1 Search Direction $q = f(x_{k-1})$

First the choice $q = f(x_{k-1})$ is investigated.

Theorem 6.3 *If $\mu > 0$ and $\nu > 0$ exist such that for any $z \in \mathbb{R}^n$ and for all vectors ξ on the straight line $\xi = x_{k-1} + \tau f(x_{k-1})$ with $\tau = \theta\tau^*$, where $0 < \theta < 1$, and τ^* minimizes $\|f(x_{k-1} + \tau^* f(x_{k-1}))\|$,*

$$z^T f'(\xi) z \geq \mu \|z\|^2, \tag{51}$$
$$\|f'(\xi)\| \leq \nu, \tag{52}$$

then for all norm-minimizing orthogonalization methods with $q_{k-1,k} = f(x_{k-1})$

$$\|f(x_k)\| \leq \left(1 - \frac{\mu^2}{\nu^2}\right)^{\frac{1}{2}} \|f(x_{k-1})\| \tag{53}$$

is valid.

Proof. Lemma 4.1 with the particular choice $q = f(x_{k-1})$ gives

$$\begin{aligned}
\|f(x_k)\|^2 &\leq \|f(x_{k-1})\|^2 - \frac{\left(f(x_{k-1})^T f'(\xi) f(x_{k-1})\right)^2}{\|f'(\xi) f(x_{k-1})\|^2} \\
&\overset{(51),(52)}{\leq} \|f(x_{k-1})\|^2 - \frac{\mu^2 \|f(x_{k-1})\|^4}{\nu^2 \|f(x_{k-1})\|^2} \\
&= \left(1 - \frac{\mu^2}{\nu^2}\right) \|f(x_{k-1})\|^2.
\end{aligned}$$

It remains to show that $\tau < \tau^*$. The vector ξ in Theorem 6.3 comes from Lemma 4.1 and is defined as $\xi = x_{k-1} + \tau f(x_{k-1})$ with $\tau = \theta\tau^*$, $0 < \theta < 1$ and τ^* minimizes $\|f(x_{k-1} + \tau^* q)\|$. □

The content of Theorem 6.3 is that norm-minimizing orthogonalization methods with $q_{k-1,k} = f(x_{k-1})$ achieve arbitrary small function values if the Jacobian is positive real (51) and continuous (52) in a neighborhood of each approximation.

Applying Lemma 5.1 to Theorem 6.3 we obtain an estimate for the errors.

Corollary 6.4 *(local convergence) $\mu > 0$ and $\nu > 0$ exist such that for any $z \in \mathbb{R}^n$ and for all vectors ξ on the straight lines $\xi = x_{i-1} + \tau_{i-1} f(x_{i-1})$ $(i = 1, \ldots, k)$ with $\tau_{i-1} = \theta\tau_{i-1}^*$, where $0 < \theta < 1$, and τ_{i-1}^* minimizes $\left\| f\left(x_{i-1} + \tau_{i-1}^* f(x_{i-1})\right) \right\|$,*

$$z^T f'(\xi) z \geq \mu \|z\|^2, \tag{54}$$

$$\|f'(\xi)\| \leq \nu. \tag{55}$$

Inequality (55) also holds for all ξ on the straight line $\xi = x + \theta(x_0 - x)$ and inequality (54) for all ξ on the straight line $\xi = x + \theta(x_k - x)$ with $0 < \theta < 1$. Then for all norm-minimizing orthogonalization methods with $q_{k-1,k} = f(x_{k-1})$

$$\|x_k - x\| \leq \frac{\nu}{\mu} \left(1 - \frac{\mu^2}{\nu^2}\right)^{\frac{k}{2}} \|x_0 - x\| \tag{56}$$

is valid.

Proof. From Lemma 5.1 we get for $\xi_i = x + \theta_i(x_i - x)$ with $0 < \theta_i < 1$ and $i = 1, k$

$$\begin{aligned}
\|x_k - x\| &\leq \|f'(\xi_0)\| \left\| [f'(\xi_k)]^{-1} \right\| \frac{\|f(x_k)\|}{\|f(x_0)\|} \|x_0 - x\| \\
&\overset{\text{Theorem 6.3}}{\leq} \|f'(\xi_0)\| \left\| [f'(\xi_k)]^{-1} \right\| \left(1 - \frac{\mu^2}{\nu^2}\right)^{\frac{k}{2}} \|x_0 - x\| \\
&\overset{(55),\,(43)}{\leq} \frac{\nu}{\mu} \left(1 - \frac{\mu^2}{\nu^2}\right)^{\frac{k}{2}} \|x_0 - x\|. \quad \square
\end{aligned}$$

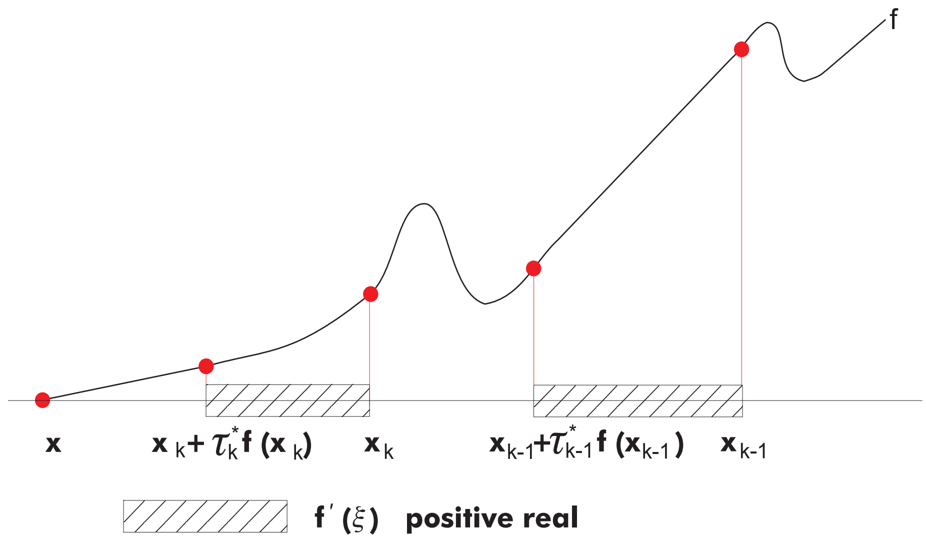

Figure 1: Sketch of the function

Corollary 6.4 assures that norm-minimizing orthogonalization methods with $q_{k-1,k} = f(x_{k-1})$ converge if the Jacobian is positive real (54) and continuous (55) in a neighborhood of each approximation. More precisely, this neighborhood is defined by the pathes from x_{i-1} to $x_{i-1} + \tau f(x_{i-1})$. The Jacobian has to be continuous on the connection between x and x_0 and positive definite on the connection between x and x_k. In Figure 1 the function is sketched with respect to the neighborhoods. Local minima can be overcome by the other search directions. The requirement of a continuous or bounded Jacobian is for the example (3) and for other elliptic or parabolic PDEs always fulfilled. Thus in this context this requirement is rather weak. The need of positive real Jacobians is harder. In example (3) this requirement is fulfilled for low Reynolds numbers.

If the Jacobian is positive real and continuous for all arguments, then the methods converge globally. Corollary 6.4 is the nonlinear equivalent to convergence estimates for the linear GMRES solver (see [14]).

6.2 Search Direction $q = [f'(x_{k-1})]^T f(x_{k-1})$

Theorem 6.5 *$\mu > 0$, $\nu > 0$ and $\omega > 0$ exist such that for any $z \in \mathbb{R}^n$ and for all vectors ξ on the straight line $\xi = x_{k-1} + \tau [f'(x_{k-1})]^T f(x_{k-1})$ with $\tau = \theta\tau^*$, where $0 < \theta < 1$, and τ^* minimizes $\left\| f\left(x_{k-1} + \tau^* [f'(x_{k-1})]^T f(x_{k-1})\right)\right\|$,*

$$z^T f'(\xi) z \geq \mu \|z\|^2, \tag{57}$$

$$\|f'(\xi)\| \leq \nu, \tag{58}$$

$$\| [f''(\xi)z]\, z\| \leq \omega \|z\|^2. \tag{59}$$

Then for all norm-minimizing orthogonalization methods with the search direction $q_{k-1,k} = [f'(x_{k-1})]^T f(x_{k-1})$

$$\|f(x_k)\| \le \left(1 - \frac{1}{\left(1 + \frac{\|f(x_{k-1})\|\omega}{\mu^2}\right)^2} \frac{\mu^2}{\nu^2}\right)^{\frac{1}{2}} \|f(x_{k-1})\| \qquad (60)$$

is valid.

Proof. For the sake of simple notations we omit the argument x_{k-1} so that $f = f(x_{k-1})$. The following inequalities hold by Lemma 6.2 with the special search direction $q = [f'(x_{k-1})]^T f$

$$\begin{aligned}
\|f(x_k)\|^2 \quad & \le \quad \|f\|^2 - \frac{1}{\left(1 + \frac{\|f\|\omega}{\mu^2}\right)^2} \frac{\left(f^T f'(x_{k-1}) \left[f'(x_{k-1})\right]^T f\right)^2}{\left\| f'(\xi) \left[f'(x_{k-1})\right]^T f \right\|^2} \\
& = \quad \|f\|^2 - \frac{1}{\left(1 + \frac{\|f\|\omega}{\mu^2}\right)^2} \frac{\left\| \left[f'(x_{k-1})\right]^T f \right\|^4}{\left\| f'(\xi) \left[f'(x_{k-1})\right]^T f \right\|^2} \\
& \le \quad \|f\|^2 - \frac{1}{\left(1 + \frac{\|f\|\omega}{\mu^2}\right)^2} \frac{\left\| \left[f'(x_{k-1})\right]^T f \right\|^2}{\|f\|^2} \|f\|^2 \frac{1}{\|f'(\xi)\|^2} \\
& \overset{(46),(58)}{\le} \quad \|f\|^2 - \frac{1}{\left(1 + \frac{\|f\|\omega}{\mu^2}\right)^2} \frac{\mu^2}{\nu^2} \|f\|^2.
\end{aligned}$$

The last inequality follows because f' is positive real by (57) and inequality (46) can be applied. □

The requirements for the Jacobian are similar to the requirements for the search direction $q = f(x_{k-1})$. In both cases continuous and positive definite Jacobians are needed. The additional requirement for a bounded or continuous second derivative, here, is weak. It is usually fulfilled for many PDEs. We can consider Figure 1 also as sketch for this search direction.

The equivalent linear CG methods (CGNE and CGNR, see [14]) are based on the normal equations and converge without the requirement of positive definiteness. Here is an essential difference to nonlinear methods. The Jacobian f' is zero for local minima and, thus, the search direction becomes zero in a local minimum. In this case the process gets stuck or only can be continued by the help of other search directions. This is the reason for the requirement of positive definiteness. Then there are no local minima. In the linear case local minima do not occur and, therefore, no additional condition is needed.

The convergence estimate (53) is better than the estimate (60) because the factor $\left(1 - \frac{\mu^2}{\nu^2}\right)^{\frac{1}{2}}$ in (53) is smaller than the factor $\left(1 - \frac{1}{\left(1 + \frac{\|f(x_{k-1})\|\omega}{\mu^2}\right)^2} \frac{\mu^2}{\nu^2}\right)^{\frac{1}{2}}$

in (60). However, for small function values the latter factor becomes close to the first one.

Similar as for the search direction $q = f(x_{k-1})$ the convergence estimate for the function values can be transferred to the errors as done in Corollary 6.4.

6.3 Search Direction $q = f'(x_{k-1})f(x_{k-1})$

Theorem 6.6 *$\mu > 0$, $\nu > 0$ and $\omega > 0$ exist such that for any $z \in \mathbb{R}^n$ and for all vectors ξ on the straight line $\xi = x_{k-1} + \tau f'(x_{k-1})f(x_{k-1})$ with $\tau = \theta\tau^*$, where $0 < \theta < 1$, and τ^* minimizes $\|f\,(x_{k-1} + \tau^* f'(x_{k-1})f(x_{k-1}))\|$,*

$$z^T f'(\xi)z \geq \mu\|z\|^2, \tag{61}$$

$$\|f'(\xi)\| \leq \nu, \tag{62}$$

$$\|\,[f''(\xi)z]\,z\| \leq \omega\|z\|^2. \tag{63}$$

Let S be the skew-symmetric part of $f'(x_{k-1})$ and

$$z^T S^2 z \geq \rho^2\|z\|^2 \tag{64}$$

with $\rho \in \mathbb{R}$. Then for all norm-minimizing orthogonalization methods with the search direction $q_{k-1,k} = f'(x_{k-1})f(x_{k-1})$

$$\|f(x_k)\| \leq \left(1 - \frac{1}{\left(1 + \frac{\|f(x_{k-1})\|\omega}{\mu^2}\right)^2}\frac{\left(\mu^2+\rho^2\right)^2}{\nu^4}\right)^{\frac{1}{2}} \|f(x_{k-1})\| \tag{65}$$

is valid.

Proof. For the sake of simple notations we omit the argument x_{k-1} so that $f = f(x_{k-1})$ and $f' = f'(x_{k-1})$. Let M be the symmetric part of $f'(x_{k-1})$. The following inequalities hold by Lemma 6.2 with the special search direction $q = f'f$

$$\begin{aligned}
\|f(x_k)\|^2 &\leq \|f\|^2 - \frac{1}{\left(1+\frac{\|f\|\omega}{\mu^2}\right)^2}\frac{\left(f^T f' f' f\right)^2}{\|f'(\xi)f'f\|^2} \\
&\stackrel{(41)}{=} \|f\|^2 - \frac{1}{\left(1+\frac{\|f\|\omega}{\mu^2}\right)^2}\frac{\left(f^T M^2 f + f^T S^2 f\right)^2}{\|f'(\xi)f'f\|^2} \\
&= \|f\|^2 - \frac{1}{\left(1+\frac{\|f\|\omega}{\mu^2}\right)^2}\left(\frac{f^T M^2 f}{\|f\|^2} + \frac{f^T S^2 f}{\|f\|^2}\right)^2 \frac{\|f\|^4}{\|f'(\xi)f'f\|^2} \\
&\stackrel{(61),(45),(64)}{\leq} \|f\|^2 - \frac{1}{\left(1+\frac{\|f\|\omega}{\mu^2}\right)^2}\left(\mu^2+\rho^2\right)^2 \frac{\|f\|^4}{\|f'(\xi)f'f\|^2}
\end{aligned}$$

$$
\begin{aligned}
&= \quad \|f\|^2 - \frac{1}{\left(1+\frac{\|f\|_\omega}{\mu^2}\right)^2}\left(\mu^2+\rho^2\right)^2 \frac{\|f'f\|^2}{\|f'(\xi)f'f\|^2}\frac{\|f\|^2}{\|f'f\|^2}\|f\|^2 \\
&\overset{(62)}{\leq} \quad \|f\|^2 - \frac{1}{\left(1+\frac{\|f\|_\omega}{\mu^2}\right)^2}\left(\mu^2+\rho^2\right)^2 \frac{1}{\nu^4}\|f\|^2 \quad \Box
\end{aligned}
$$

The requirements for the Jacobian are similar to the requirements for the above investigated search directions. In all cases continuous and positive definite Jacobians are needed. Again Figure 1 is a sketch for this search direction.

But the convergence estimate (65) is different to the latter two cases. If ρ is small, then the part $\frac{\mu^4}{\nu^4}$ dominates (65) and leads to an estimate essentially slower decreasing than the corresponding estimates of the other two search directions with the factor $\frac{\mu^2}{\nu^2}$. On the other hand, if we consider equation (3), then $\frac{\mu}{\nu} \rightarrow 0$ for an increasing Reynolds number R and the estimate becomes worse and worse for the search directions $q = f(x_{k-1})$ and $q = [f'(x_{k-1})]^T f(x_{k-1})$. On the contrary, in this case we get a quite different estimate for the search direction $q = f'(x_{k-1})f(x_{k-1})$. Let $M(\xi)$ be the symmetric part of $f'(\xi)$ and $S(\xi)$ be the skew-symmetric part of $f'(\xi)$ and let us assume that $\|M(\xi)\| \leq \Lambda$ and $\|S(\xi)\| \leq \Phi$. If R in (3) increases, then the skew symmetric part of the Jacobian increases. Let us assume that the values ρ and Φ of the skew symmetric part increase in the same way.

$$
\begin{aligned}
\frac{\mu^2+\rho^2}{\nu^2} &= \frac{\mu^2+\rho^2}{\|f'(\xi)\|^2} = \frac{\mu^2+\rho^2}{\|M(\xi)+S(\xi)\|^2} \\
&\geq \frac{\mu^2+\rho^2}{\Lambda^2+2\Lambda\Phi+\Phi^2} \\
&= \frac{\frac{\mu^2}{\rho^2}+1}{\frac{\Lambda^2}{\rho^2}+2\frac{\Lambda\Phi}{\rho^2}+\frac{\Phi^2}{\rho^2}} \rightarrow \frac{1}{c^2}
\end{aligned}
$$

if $\rho \rightarrow \infty$, $\Phi \rightarrow \infty$ and $\frac{\Phi}{\rho} \rightarrow c$. Thus, we get in (65) convergence for high Reynolds numbers.

Similar as for the search direction $q = f(x_{k-1})$ the convergence estimate for the function values can be transferred to the errors as done in Corollary 6.4.

7 Numerical Experiments

As test example the partial differential equation (3) was discretized by second order finite differences on a rectangular domain with Dirichlet boundary conditions. The function h in (3) was chosen in such a way that prescribed trigonometric functions are the true solution of the discretized problem. Thus, the exact errors could be calculated in addition to the function values. Note that the Jacobian is positive definite for small values of R but becomes indefinite for larger values.

We tested as reference method the classical Newton linearisation in connection with modern iterative solvers. The linear solver was chosen adaptively according to the polyalgorithm strategy [13]. Thus the nearly optimal method is chosen from the three methods PRES20, BICO and ATPRES (see [14]). As new orthogonalization method we take a truncated method with

$$\tilde{x}_k = x_{k-1},$$
$$q_{k-1} = f(x_{k-1}),$$
$$q_{k-i-1} = x_{k-i}$$

for $i = 1, \ldots, 4$. This method is denoted by F5. The method FDF5 has the additional search direction

$$q_{k-6} = f'(x_{k-2})f(x_{k-2}).$$

This search direction is not the theoretical investigated $f'(x_{k-1})f(x_{k-1})$ but close. It is this direction delayed by one iteration step. The reason for taking this modified search direction is that it is availabel without additional computational costs from the previous steps. The nonlinear system of dimension 5, 6 respectively, was approximated by one step of Newton's method and a direct Gaussian elimination.

The convergence history is depicted in figure 2 for $R = 1$ and in figure 2 for $R = 50$. For $R = 100$ none of the methods converges and for this case the Jacobian is indefinite so that the here introduced theory does not guarantee convergence. As iteration for the Newton method we count the combination of inner and outer iterations according to formula (11). Thus the computational work for one "iteration" of the classical Newton is one matrix-vector multiplication and several dot products depending on the chosen linear solver. For the two new methods we need a function evaluation in each iteration step. Note that the partial derivatives are obtained without additional computational costs as auxiliary values during the function evaluation. In addition a cheap Gaussian elimination for a system of dimension 5 or 6 is necessary. All three methods require the same amount of triadic operations (vector = vector + scalar $\times$ vector). Thus the real computational amount corresponds to the number of iterations (for the classical approach the highly tuned commercial code FIDISOL [13] was used while for the new method a simple test program was developed. Thus the real cpu times cannot be compared).

For both Reynolds numbers the two new methods converge essentially faster. Moreover, the speed of convergence does not essentially depend on the Reynolds number for the new methods as predicted theoretically while the convergence of Newtons method slows down with increasing R.

8 Discussion

All three investigated search directions need positive definite Jacobians and the estimates do not guarantee convergence if there are local minima. The reason

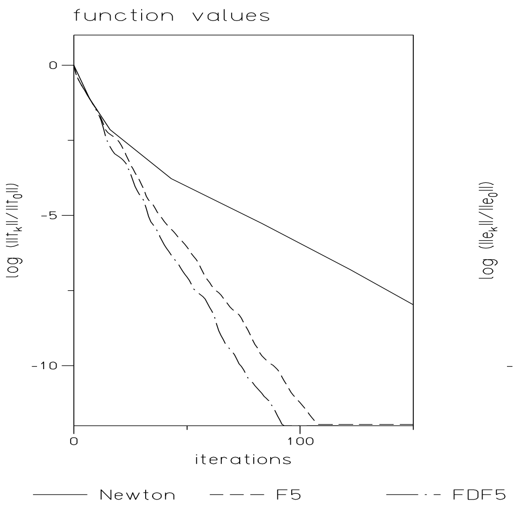

Figure 2: Sample problem with R=1

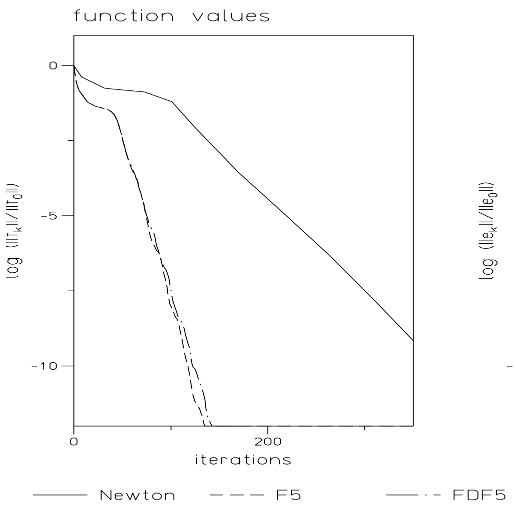

Figure 3: Sample problem with R=50

for this is that we have chosen a norm-minimizing approach. The iterates will get stuck in local minima. In such cases non-minimizing approaches may be succesful. However, the investigation of these approaches was postponed for performance reasons and is beyond the scope of this article.

Looking at the convergence estimates suggests to choose a combination of the two search directions $q = f(x_{k-1})$ and $q = f'(x_{k-1})f(x_{k-1})$. The first one assures convergence for postive definite Jacobians with a small skew-symmetric part while the second one gives a better convergence estimate for larger skew-symmetric parts.

In this article theoretical aspects have been considered. The theorems give conditions for convergence. The numerical experiments confirm the theoretical results. Moreover, two new methods have been derived canonicly from the theoretical framework that seem to be very promising. In the here performed tests these methods are faster by a factor of two compared to classical approaches.

References

[1] O. Axelsson and A. T. Chronopoulos. On nonlinear generalized conjugate gradient methods. *Numer. Math.*, 69:1–15, 1994.

[2] C. G. Broyden. A class of methods for solving nonlinear simultaneous equations. *Mathematics of Computation*, 19(138):577–593, 1965.

[3] A. T. Chronopoulos. Nonlinear CG-like iterative methods. Research Report UMSI 91/99, University of Minnesota, Supercomputer Institute, 1200 Washington Avenue South, Minneapolis, Minnesota 55415, USA, 1991.

[4] J. W. Daniel. Convergence of the conjugate gradient method with computational convenient modifications. *Numerische Mathematik*, 10:125–131, 1967.

[5] A. Edelman and S. T. Smith. On conjugate gradient-like methods for eigen-like problems. In L. Adams and J. L. Nazareth, editors, *Linear and Nonlinear Conjugate Gradient-Related Methods*, pages 24–36. SIAM, 1995.

[6] R. Fletcher and M. J. D. Powell. A rapidly convergent descent method for minimization. *Computer Journal*, 6:163–168, 1963.

[7] R. Fletcher and C. M. Reeves. Function minimization by conjugate gradients. *Computer Journal*, 7:149–154, 1964.

[8] J. C. Gilbert and J. Nocedal. Global convergence properties of conjugate gradient methods for optimization. *SIAM Journal on Optimization*, 2:21–42, 1992.

[9] A. Griewank. Direct calculation Newton steps without accumulating Jacobians. In T. F. Coleman and Li Yuying, editors, *Large-Scale Numerical Optimization*, pages 115–137. SIAM, 1990.

[10] J. Nocedal. Theory of algorithms for unconstrained optimization. *Acta Numerica*, pages 199–242, 1992.

[11] J. Nocedal. Conjugate gradient methods and nonlinear optimization. In L. Adams and J. L. Nazareth, editors, *Linear and Nonlinear Conjugate Gradient-Related Methods*, pages 9–23. SIAM, 1995.

[12] E. Polak and G. Ribière. Note sur la convergence de méthodes de directions conjuguées. *Revue Française d'Information et de Recherche Opérationelle*, 16:35–43, 1969.

[13] W. Schönauer. *Scientific Computing on Vector Computers.* North-Holland, Amsterdam, New York, Oxford, Tokyo, 1987.

[14] R. Weiss. *Parameter-Free Iterative Linear Solvers.* Mathematical Research, vol. 97. Akademie Verlag, Berlin, 1996.