NUMERICAL METHODS FOR UNCONSTRAINED OPTIMIZATION [1]

(student text)

Ladislav Lukšan [2]

1 Introduction

In this text, basic methods for unconstrained minimization are studied including their convergence properties. After a short introduction, section 2 describes line search methods and their typical realizations (conjugate gradient methods and variable metric methods). Section 3 is devoted to trust region methods, which are suitable especially for globally convergent realization of the Newton method and the Gauss-Newton method for minimization of sum of squares. Special methods for large-scale sparse problems are described in section 4 . Section 5 is devoted to methods for solving systems of nonlinear equations. Special methods for solving large-scale sparse systems of nonlinear equations are described in Section 6. Section 7 contains some results concerning optimization of dynamical systems. Theorems and lemmas are always proved in this text. Propositions, which can be found in standard student texts, are introduced without proofs. Most of missing proofs can be found in the book L.Lukšan: Variable Metric Methods, Academia, Prague 1990. Since most of statements are carefully proved, we do not quote references in this text. For further study of methods for unconstrained optimization we recommend the following books. T.F.Coleman: Large sparse numerical optimization, Springer-Verlag, Berlin, 1984. J.E.Dennis, R.B.Schnabel: Numerical methods for unconstrained optimization and nonlinear equations, Prentice-Hall, Englewood Cliffs, New Jersey, 1983. R.Fletcher: Practical methods of optimization, Wiley, New York, 1987. P.E.Gill, W.Murray, M.H.Wright: Practical optimization, Academic Press, London, 1981. C.T.Kelley: Iterative Methods for Linear and Nonlinear Equations, SIAM, Philadelphia, Pennsylvania, 1995. G.H.Golub, C.F.Van Loan: Matrix computations, Johns Hopkins University Press, Baltimore, 1989. M.M.Mäkelä, J.Neittaanmäki: Nonsmooth Optimization. World Scientific Publishing Co. Ltd., London, 1992.

1.1 Basic notions and notations

We will use the notation:

$$x \in R^n \text{ for } n-\text{dimensional vector},$$

$$F(x) \text{ for function } F: R^n \to R,$$

$$g(x) = [\partial F/\partial x_1, \ldots, \partial F/\partial x_n]^T,$$

$$G(x) = \begin{bmatrix} \partial^2 F/\partial x_1^2, & \ldots, & \partial^2 F/\partial x_1 \partial x_n \\ \vdots & \vdots & \vdots \\ \partial^2 F/\partial x_n \partial x_1, & \ldots, & \partial^2 F/\partial x_n^2 \end{bmatrix}.$$

Here $F(x)$ is the objective function, $g(x)$ is its gradient and $G(x)$ is its Hessian matrix (matrix of second order partial derivatives). Continuity of second order partial derivatives implies symmetry of the matrix $G(x)$.

[1] This work was supported by grant GA ČR No. 201/96/0918 and by CNR, GNIM visiting professorship program
[2] Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic (luksan@uivt.cas.cz)

For investigation of convergence of optimization methods, we will often use the assumptions (F1)-(F5):

Definition 1 We say that the function $F : R^n \to R$ is bounded from below, if

$$F(x) \geq \underline{F} \qquad \forall x \in R^n. \tag{F1}$$

Definition 2 We say that the function $F : R^n \to R$ has compact levels, if the set

$$\mathcal{L}(\overline{F}) = \{x \in R^n : F(x) \leq \overline{F}\} \tag{F2}$$

is compact $\forall \overline{F}$ (the empty set is supposed to be compact).

Definition 3 We say that the function $F : R^n \to R$ has bounded second order derivatives if

$$|d^T G(x) d| \leq \overline{G} \parallel d \parallel^2 \tag{F3}$$

$\forall x \in R^n$, $\forall d \in R^n$. This is equivalent to the condition $\parallel G(x) \parallel \leq \overline{G}$ $\forall x \in R^n$.

Remark 1 Instead of boundedness of the second order derivatives, we often use the Lipschitz continuity of the first derivatives:

$$\parallel g(x+d) - g(x) \parallel \leq \overline{G} \parallel d \parallel$$

$\forall x \in R^n$, $\forall d \in R^n$

Definition 4 We say that the function $F : R^n \to R$ is uniformly (or strongly) convex, if

$$d^T G(x) d \geq \underline{G} \parallel d \parallel^2 \tag{F4}$$

$\forall x \in R^n$, $\forall d \in R^n$ (here $\underline{G} > 0$).

Definition 5 We say that the function $F : R^n \to R$ has Lipschitz continuous second order derivatives, if

$$\parallel G(x+d) - G(x) \parallel \leq \overline{L} \parallel d \parallel \tag{F5}$$

$\forall x \in R^n$, $\forall d \in R^n$.

In proofs of convergence, we will often use mean value theorems:

(**a**) $F(x+d) = F(x) + d^T g(x) + \frac{1}{2} d^T G(\tilde{x}) d$, where $\tilde{x} = x + \tilde{\lambda} d$ a $0 \leq \tilde{\lambda} \leq 1$.

From (F3) we get

$$F(x+d) - F(x) \leq d^T g(x) + \frac{1}{2} \overline{G} \parallel d \parallel^2 .$$

From (F4) we get

$$F(x+d) - F(x) \geq d^T g(x) + \frac{1}{2} \underline{G} \parallel d \parallel^2 .$$

(**b**) $g(x+d) = g(x) + \int_0^1 G(x + \lambda d) d d\lambda$.

From (F3) we obtain

$$\parallel g(x+d) - g(x) \parallel \leq \overline{G} \parallel d \parallel,$$
$$d^T (g(x+d) - g(x)) \leq \overline{G} \parallel d \parallel^2 .$$

From (F4) we obtain

$$\| g(x+d) - g(x) \| \geq \underline{G} \| d \|,$$
$$d^T(g(x+d) - g(x)) \geq \underline{G} \| d \|^2 .$$

Proof of the last proposition:

$$d^T(g(x+d) - g(x)) = \int_0^1 d^T G(x+\lambda d) d d\lambda \geq \int_0^1 \underline{G} \| d \|^2 d\lambda = \underline{G} \| d \|^2,$$

$$\underline{G} \| d \|^2 \leq d^T(g(x+d) - g(x)) \leq \| d \| \| g(x+d) - g(x) \| .$$

1.2 Optimality conditions

Definition 6 We say that the point $x^* \in R^n$ is a local minimizer of the function $F : R^n \to R$ if a number $\varepsilon > 0$ exists, such that

$$F(x^*) \leq F(x) \qquad \forall x \in \mathcal{B}(x^*, \varepsilon),$$

where $\mathcal{B}(x^*, \varepsilon) = \{x \in R^n : \| x - x^* \| < \varepsilon\}$. If in addition $F(x^*) < F(x)$ when $x^* \neq x$, then we say that the point $x^* \in R^n$ is a strict local minimizer of the function $F : R^n \to R$.

Proposition 1 Let the point $x^* \in R^n$ be a local minimizer of the function $F : R^n \to R$ and let $F \in C^1$ (continuously differentiable) on $\mathcal{B}(x^*, \varepsilon)$. Then

$$g(x^*) = 0.$$

If in addition $F \in C^2$ (two times continuously differentiable) on $\mathcal{B}(x^*, \varepsilon)$, then

$$G(x^*) \geq 0$$

(the matrix $G(x^*)$ is positive semidefinite)

Proposition 2 Let $F : R^n \to R \in C^2$ on $\mathcal{B}(x^*, \varepsilon)$ and let

$$g(x^*) = 0$$

and

$$G(x^*) > 0$$

(the matrix $G(x^*)$ is positive definite). Then the point $x^* \in R^n$ is a strict local minimizer of the function $F : R^n \to R$.

1.3 Basic notions from the theory of convergence

Now we will investigate properties of convergent sequences.

Definition 7 Let $x_i \in R^n$, $i \in N$, be a sequence of points. If for an arbitrary $\varepsilon > 0$ an index $k \in N$ exists, such that $x_i \in \mathcal{B}(x^*, \varepsilon)$ $\forall i \geq k$, then we say that the sequence $x_i \in R^n$, $i \in N$ converges to the point $x^* \in R^n$ and we write $x_i \to x^*$. We use the notation $F_i = F(x_i)$, $g_i = g(x_i)$, $G_i = G(x_i)$.

Theorem 1 Let $x_i \in R^n$ and $d_i \in R^n$, $i \in N$, be two sequences. Let $e_i = x_i - x^*$, $i \in N$, where $x^* \in R^n$ is a local minimizer of the function $F \in C^2$, which satisfy the condition (F5). Then

$$F(x_i + d_i) - F(x_i) = d_i^T g_i + \frac{1}{2} d_i^T G_i d_i + O(\| d_i \|^3),$$

$$g(x_i + d_i) - g(x_i) = G_i d_i + O(\| d_i \|^2)$$

and

$$F(x_i) - F(x^*) = \frac{1}{2} e_i^T G^* e_i + O(\| e_i \|^3),$$

$$g(x_i) = G^* e_i + O(\| e_i \|^2).$$

Here $\| O(\xi_i) \| \leq C \| \xi_i \|$ if $\| \xi_i \| \rightarrow 0$.

Proof From (F5) and from the mean value theorem (a), it follows

$$\begin{aligned} F(x_i + d_i) - F(x_i) &= d_i^T g_i + \frac{1}{2} d_i^T G(x_i + \tilde{\lambda} d_i) d_i = \\ &= d_i^T g_i + \frac{1}{2} d_i^T G_i d_i + \frac{1}{2} d_i^T (G(x_i + \tilde{\lambda} d_i) - G_i) d_i \leq \\ &\leq d_i^T g_i + \frac{1}{2} d_i^T G_i d_i + \frac{1}{2} \| G(x_i + \tilde{\lambda} d_i) - G_i \| \| d_i \|^2 \leq \\ &\leq d_i^T g_i + \frac{1}{2} d_i^T G_i d_i + \frac{\tilde{\lambda}}{2} \overline{L} \| d_i \|^3, \end{aligned}$$

where $0 \leq \tilde{\lambda} \leq 1$. Similarly from (F5) and from the mean value theorem (b), it follows

$$\begin{aligned} g(x_i + d_i) - g(x_i) &= \int_0^1 G(x_i + \lambda d_i) d_i d\lambda = \\ &= G_i d_i + \int_0^1 (G(x_i + \lambda d_i) - G_i) d_i d\lambda \end{aligned}$$

and

$$\begin{aligned} \| \int_0^1 (G(x_i + \lambda d_i) - G_i) d_i d\lambda \| &\leq \int_0^1 \| G(x_i + \lambda d_i) - G_i \| \| d_i \| d\lambda \leq \\ &\leq \overline{L} \int_0^1 \| d_i \|^2 d\lambda = \overline{L} \| d_i \|^2 . \end{aligned}$$

Hence we have proved the first two relations. The second two relations can be proved by the same way if we use x_i instead of $x_i + d_i$, x^* instead of x_i, $e_i = x_i - x^*$ instead of $d_i = x_i + d_i - x_i$ and take into account that $g(x^*) = 0$.

Remark 2 Without using the condition (F5), we can derive the following estimations

$$F(x_i + d_i) - F(x_i) = d_i^T g_i + \frac{1}{2} d_i^T G_i d_i + o(\| d_i \|^2),$$

$$g(x_i + d_i) - g(x_i) = G_i d_i + o(\| d_i \|)$$

and

$$F(x_i) - F(x^*) = \frac{1}{2} e_i^T G^* e_i + o(\| e_i \|^2),$$

$$g(x_i) = G^* e_i + o(\| e_i \|)$$

Here $\| o(\xi_i) \| / \| \xi_i \| \rightarrow 0$ if $\| \xi_i \| \rightarrow 0$.

Definition 8 Let $x_i \rightarrow x^*$. If an index $k \in N$ and values $0 < M_k < \infty$ and $0 < q < 1$ exist, so that

$$\| x_i - x^* \| \leq M_k q^{i-k} \| x_k - x^* \|$$

$\forall i \geq k$, we say that the sequence $x_i \in R^n$, $i \in N$, converges to the point x^* with the R-linear rate.

Theorem 2 The sequence $x_i \in R^n$, $i \in N$, converges to the point x^* with the R-linear rate, if and only if

$$\limsup_{i\to\infty} \| x_i - x^* \|^{\frac{1}{i}} = \hat{q} < 1$$

($\limsup$ exists and it is less that 1).

Proof Definition 8 implies

$$\| x_i - x^* \| \leq M_k q^{i-k} \| x_k - x^* \|$$

$\forall i \geq k$, where $q < 1$, so that

$$\| x_i - x^* \|^{\frac{1}{i}} \leq (M_k \| x_k - x^* \|)^{\frac{1}{i}} q^{1-\frac{k}{i}} \leq \max(1, M_k \| x_k - x^* \|).$$

This sequence is bounded, so that $\limsup$ exists and

$$\limsup_{i\to\infty} \| x_i - x^* \|^{\frac{1}{i}} = \hat{q} \leq \lim_{i\to\infty} (M_k \| x_k - x^* \|)^{\frac{1}{i}} q^{1-\frac{k}{i}} = q < 1.$$

On the other hand, let the expression on the left hand side of the last inequality exist and be equal to $\hat{q} < 1$. Then, for an arbitrary number $\hat{q} < q < 1$, an index $k \in N$ exists, so that

$$\| x_i - x^* \|^{\frac{1}{i}} \leq q$$

or

$$\| x_i - x^* \| \leq q^i$$

$\forall i \geq k$. Choose

$$M_k = \frac{q^k}{\| x_k - x^* \|}.$$

Then

$$\| x_i - x^* \| \leq M_k q^{i-k} \| x_k - x^* \|$$

Definition 9 The sequence $x_i \in R^n$, $i \in N$, converges to the point x^* with the R-superlinear rate, if

$$\lim_{i\to\infty} \| x_i - x^* \|^{\frac{1}{i}} = 0.$$

Definition 10 The sequence $x_i \in R^n$, $i \in N$, converges to the point x^* with Q-linear rate, if an index $k \in N$ and a constant $0 < q < 1$ exist, so that

$$\frac{\| x_{i+1} - x^* \|}{\| x_i - x^* \|} \leq q \qquad \forall i \geq k.$$

Definition 11 The sequence $x_i \in R^n$, $i \in N$, converges to the point x^* with Q-superlinear rate, if

$$\lim_{i\to\infty} \frac{\| x_{i+1} - x^* \|}{\| x_i - x^* \|} = 0$$

Theorem 3 Let $x_i \to x^*$ with Q-linear rate (Q-superlinear rate). Then $x_i \to x^*$ with R-linear rate (R-superlinear rate).

Proof The R-linear convergence follows from the Q-linear convergence immediately (it suffices to choose $M_k = 1$). Let $0 < q < 1$ be an arbitrary (small) number. From the Q-superlinear convergence, existence of an index $k \in N$ follows, such that

$$\frac{\| x_{i+1} - x^* \|}{\| x_i - x^* \|} \leq q \qquad \forall i \geq k,$$

so that by Theorem 2 it holds

$$\limsup_{i \to \infty} \| x_i - x^* \|^{\frac{1}{i}} \leq q.$$

Since q is an arbitrary (small) number, it follows that

$$\lim_{i \to \infty} \| x_i - x^* \|^{\frac{1}{i}} = 0.$$

Remark 3 The Q-linear (Q-superlinear) convergence implies monotonicity of the sequence $\| x_i - x^* \|$, $i \in N$ (starting with a suitable index $k \in N$).

Definition 12 The sequence $x_i \in R^n$, $i \in N$, converges to the point x^* with Q-quadratic rate, if an index $k \in N$ and a constant $0 < M_k < \infty$ exist, so that

$$\| x_{i+1} - x^* \| \leq M_k \| x_i - x^* \|^2 \qquad \forall i \geq k.$$

Definition 13 The sequence $x_i \in R^n$, $i \in N$, converges to the point x^* with Q-m-step quadratic rate, if an index $k \in N$, a number $m \in N$ and a constant $0 < M_k < \infty$ exist, so that

$$\| x_{i+m} - x^* \| \leq M_k \| x_i - x^* \|^2 \qquad \forall i \geq k$$

Remark 4 Sometimes we use the weaker assumption $\| x_{jl+m} - x^* \| \leq M \| x_{jl} - x^* \|^2$ $\forall j \in N$ (selected sequence $i = jl$).

1.4 Basic optimization methods

The basic optimization method is an iterative process generating the sequence $x_i \in R^n$, $i \in N$, such that

$$x_{i+1} = x_i + \alpha_i s_i,$$

where the direction vector $s_i \in R^n$ is determined from the values x_j, F_j, g_j, G_j, $1 \leq j \leq i$, and the stepsize $\alpha_i > 0$ is determined from the behaviour of the function $F : R^n \to R$ in the neighborhood of the point $x_i \in R^n$.

Definition 14 We say that the basic optimization method is globally convergent, if

$$\liminf_{i \to \infty} \| g(x_i) \| = 0$$

holds for an arbitrary initial point $x_1 \in R^n$. The best known optimization methods are the steepest descent method and the Newton method.

The steepest descent method is defined by the formulas

$$s_i = -g(x_i),$$

$$\alpha_i = \arg\min_{\alpha \geq 0} F(x_i + \alpha s_i).$$

Advantages:

- It is globally convergent.
- It uses only vectors from R^n.

  $O(n)$ - storage elements.

  $O(n)$ - operations per iteration.

Disadvantages:

- It requires perfect line search.
- it is only R-linearly convergent with the asymptotic rate

$$\limsup_{i\to\infty} \| x_i - x^* \|^{\frac{1}{i}} \le \frac{\kappa(G(x^*)) - 1}{\kappa(G(x^*)) + 1}.$$

  This estimate of the asymptotic rate is usually realistic (not overestimated). For instance, if $\kappa(G(x^*)) = 10^3$, then we require approximately 4600 iterations for decreasing the error $\| x - x^* \|$ about 4 orders. If $\kappa(G(x^*)) = 10^6$ then we require approximately 9200000 iterations for decreasing the error $\| x - x^* \|$ about 8 orders.

The Newton method is defined by the formulas

$$s_i = -G^{-1}(x_i)g(x_i),$$

$$\alpha_i = 1.$$

Advantages:

- It is Q-quadratically convergent. If it converges, then the local minimizer is found after few iterations.
- It uses a simple line search.

Disadvantages:

- It is not globally convergent. If x_1 is far from x^*, then it may not converge.
- It uses a matrix and it is necessary to solve a system of linear equations.

  $O(n^2)$ - storage elements.

  $O(n^3)$ - operations per iteration.
- It is necessary to compute second order derivatives.

To remove disadvantages of these simple methods, more efficient but more complicated methods has been developed.

a) Line search methods developed from the steepest descent method:

- They use imperfect line search.
- They possess higher rate of convergence (conjugate direction principle).

b) Trust region methods developed from the Newton method:

- They have guaranteed global convergence.
- They require a lower number of operations (inexact solution of linear equations).

2 Line search methods

2.1 Basic properties of the line search methods

The definition of descent directions and uniformly descent directions is crucial for the construction of line search methods.

Definition 15 We say that the direction vectors $s_i \in R^n$, $i \in N$, are descent if

$$s_i^T g_i < 0 \tag{S1a}$$

$\forall i \in N$. We say that the direction vectors $s_i \in R^n$, $i \in N$ are uniformly descent, if a constant $0 < \varepsilon_0 \leq 1$ exists, such that

$$-s_i^T g_i \geq \varepsilon_0 \parallel s_i \parallel \parallel g_i \parallel \tag{S1b}$$

$\forall i \in N$.

The direction vectors $s_i \in R^n$, $i \in N$, are frequently determined as inexact solutions of systems of linear equations $B_i s_i + g_i = 0$, $i \in N$.

Theorem 4 Let B_i, $i \in N$, be a sequence of symmetric positive definite matrices, such that $\underline{B} \leq \underline{\lambda}(B_i) \leq \overline{\lambda}(B_i) \leq \overline{B}$ $\forall i \in N$ ($\underline{\lambda}(B_i)$ and $\overline{\lambda}(B_i)$ are the least and the greatest eigenvalues of the matrix B_i, respectively). Let $\parallel B_i s_i + g_i \parallel \leq \overline{\omega} \parallel g_i \parallel$ $\forall i \in N$, where $\overline{\omega} < \underline{B}/\overline{B}$. Then the condition (S1b) is satisfied with

$$\varepsilon_0 = \frac{\underline{B}/\overline{B} - \overline{\omega}}{1 + \overline{\omega}}.$$

Proof Denote $r_i = B_i s_i + g_i$, $i \in N$, so that $\parallel r_i \parallel \leq \overline{\omega} \parallel g_i \parallel$, $i \in N$. Obviously $s_i = B_i^{-1}(r_i - g_i)$, $i \in N$, so that

$$-s_i^T g_i = g_i^T B_i^{-1} g_i - r_i^T B_i^{-1} g_i \geq \parallel g_i \parallel^2 / \overline{\lambda}(B_i) - \parallel r_i \parallel \parallel g_i \parallel / \underline{\lambda}(B_i) \geq (1/\overline{B} - \overline{\omega}/\underline{B}) \parallel g_i \parallel^2 .$$

Furthermore

$$\parallel s_i \parallel = \parallel B_i^{-1}(r_i - g_i) \parallel \leq (\parallel r_i \parallel + \parallel g_i \parallel)/\underline{\lambda}(B_i) \leq (1 + \overline{\omega}) \parallel g_i \parallel / \underline{B},$$

which after substitution gives

$$-\frac{s_i^T g_i}{\parallel s_i \parallel \parallel g_i \parallel} \geq \frac{1/\overline{B} - \overline{\omega}/\underline{B}}{(1 + \overline{\omega})/\underline{B}} = \frac{\underline{B}/\overline{B} - \overline{\omega}}{1 + \overline{\omega}} \triangleq \varepsilon_0$$

Another substantial part of line search methods is the stepsize selection, which has to satisfy some restrictive conditions.

Definition 16 We say that the stepsize $\alpha_i > 0$, $i \in N$, satisfies either a strong Wolfe condition or a weak Wolfe condition or a Goldstein condition or an Armijo condition, if constants $0 < \varepsilon_1 < 1/2$ and $\varepsilon_1 < \varepsilon_2 < 1$ (independent of the index $i \in N$) exist, so that

$$F_{i+1} - F_i \leq \varepsilon_1 \alpha_i s_i^T g_i \tag{S2}$$

and either

$$|s_i^T g_{i+1}| \leq \varepsilon_2 |s_i^T g_i| \tag{S3a}$$

or

$$s_i^T g_{i+1} \geq \varepsilon_2 s_i^T g_i \tag{S3b}$$

or

$$F_{i+1} - F_i \geq \varepsilon_2 \alpha_i s_i^T g_i \tag{S3c}$$

or $\alpha_i > 0$ is the first member of the sequence α_i^j, $j \in N$, such that $\alpha_i^1 = 1$ and

$$\underline{\beta} \alpha_i^j \leq \alpha_i^{j+1} \leq \overline{\beta} \alpha_i^j \qquad \forall j \in N, \tag{S3d}$$

for which (S2) holds ($0 < \underline{\beta} \leq \overline{\beta} < 1$ are some constant).

Conditions (S1)-(S3) form a basis for the definition of line search methods.

Definition 17 We say that the basic optimization method $x_{i+1} = x_i + \alpha_i s_i$, $i \in N$, is a descent line search method, if the direction vectors $s_i \in R^n$, $i \in N$, satisfy the condition (S1a) and the stepsizes $\alpha_i > 0$, $i \in N$, satisfy the condition (S2) and one of the conditions (S3a)-(S3d). We say that the basic optimization method $x_{i+1} = x_i + \alpha_i s_i$, $i \in N$, is a uniformly descent line search method, if it is a descent line search method and (S1b) holds.

Remark 5 Conjugate gradient methods derived from the steepest descent method utilize the condition (S3a). Variable metric methods derived from the Newton method (where $\alpha_i \to 1$ for $i \to \infty$) use the condition (S3b). Derivative free methods utilize the condition (S3c). Nonsmooth methods use the condition (S3d).

Remark 6 The steepest descent method is a uniformly descent line search method since $s_i = -g_i$, so that $s_i^T g_i = - \| g_i \|^2 = - \| s_i \| \| g_i \|$ and (S1b) holds for $\varepsilon_0 = 1$.

Lemma 1 (Consistency) Let the function $F : R^n \to R$ satisfy the conditions (F1) and (F3) and the direction vector s_i satisfy the condition (S1a). Then for any of the conditions (S3a)-(S3d) a stepsize $\alpha_i > 0$ exists, which satisfies this condition together with the condition (S2).

Proof (for (S3b)). Denote

$$\mathcal{M}_i = \{\alpha \geq 0 : F(x_i + \alpha s_i) - F_i \leq \varepsilon_1 \alpha s_i^T g_i\}.$$

Obviously $\mathcal{M}_i \neq \emptyset$ since $0 \in \mathcal{M}_i$. Let $\alpha_i = \sup \mathcal{M}_i$. By (F1) it holds that $F(x_i + \alpha s_i) \geq \underline{F}$ so that for $\alpha_i \in \mathcal{M}_i$ we obtain $\alpha_i \leq (\underline{F} - F_i)/\varepsilon_1 s_i^T g_i < \infty$. Now we show that $s_i^T g(x_i + \alpha_i s_i) \geq \varepsilon_1 s_i^T g_i$ holds. Suppose by contradiction that

$$s_i^T g(x_i + \alpha_i s_i) = \varepsilon s_i^T g_i,$$

where $\varepsilon > \varepsilon_1$. Then (F3) implies

$$\begin{aligned} F(x_i + \alpha s_i) - F_i &\leq F(x_i + \alpha_i s_i) - F_i + s_i^T g(x_i + \alpha_i s_i)(\alpha - \alpha_i) + \frac{1}{2}\overline{G} \| s_i \|^2 (\alpha - \alpha_i)^2 \leq \\ &\leq \varepsilon_1 \alpha_i s_i^T g_i + \varepsilon(\alpha - \alpha_i) s_i^T g_i + \frac{1}{2}\overline{G} \| s_i \|^2 (\alpha - \alpha_i)^2 = \\ &= \alpha \varepsilon_1 s_i^T g_i - (\varepsilon_1 - \varepsilon)(\alpha - \alpha_i) s_i^T g_i + \frac{1}{2}\overline{G} \| s_i \|^2 (\alpha - \alpha_i)^2 \end{aligned}$$

and for $\alpha = \alpha_i + (\varepsilon_1 - \varepsilon) s_i^T g_i / \overline{G} \| s_i \|^2 > \alpha_i$ we get

$$F(x_i + \alpha s_i) - F_i \leq \alpha \varepsilon_1 s_i^T g_i - \frac{1}{2} \frac{(\varepsilon - \varepsilon_1)^2 (s_i^T g_i)^2}{\overline{G} \| s_i \|^2} < \alpha \varepsilon_1 s_i^T g_i,$$

which is a contradiction since $\alpha_i = \sup \mathcal{M}_i$. From $s_i^T g(x_i + \alpha_i s_i) \geq \varepsilon_1 s_i^T g_i$ a $s_i^T g_i < 0$ it follows that $\alpha_i > 0$.

In the subsequent text, we will usually suppose that (S1b) holds, since uniformly descent line search methods have very advantageous properties.

Theorem 5 (Global convergence) Let the function $F : R^n \to R$ satisfy conditions (F1) and (F3). Then each uniformly descent line search method is globally convergent.

Proof (for (S3b)) First we prove that a constant $\underline{c} > 0$ exists, such that for an arbitrary index $i \in N$ it holds

$$\alpha_i \geq \underline{c} \parallel g_i \parallel / \parallel s_i \parallel \tag{a}$$

and

$$F_{i+1} - F_i \leq -\varepsilon_0 \varepsilon_1 \underline{c} \parallel g_i \parallel . \tag{b}$$

From the condition (S3b) we obtain

$$\varepsilon_2 s_i^T g_i \leq s_i^T g(x_i + \alpha_i s_i) \leq s_i^T g_i + \alpha_i \overline{G} \parallel s_i \parallel^2$$

or

$$\alpha_i \geq \frac{(\varepsilon_2 - 1) s_i^T g_i}{\overline{G} \parallel s_i \parallel^2} \geq \frac{\varepsilon_0 (1 - \varepsilon_2) \parallel g_i \parallel}{\overline{G} \parallel s_i \parallel},$$

so that (a) holds for $\underline{c} = \varepsilon_0 (1 - \varepsilon_2)/\overline{G} > 0$. Furthermore, from the condition (S2) and from (a) we get

$$F_{i+1} - F_i \leq \varepsilon_1 \alpha_i s_i^T g_i \leq -\varepsilon_0 \varepsilon_1 \alpha_i \parallel s_i \parallel \parallel g_i \parallel \leq -\varepsilon_0 \varepsilon_1 \underline{c} \parallel g_i \parallel^2,$$

so that (b) holds. Now we can write

$$F_{i+1} = F_1 + \sum_{j=1}^{i} (F_{j+1} - F_j) \leq F_1 - \varepsilon_0 \varepsilon_1 \underline{c} \sum_{j=1}^{i} \parallel g_j \parallel^2 .$$

The sequence F_i, $i \in N$ is decreasing by (b) and bounded from below by (F1). Therefore, the limit

$$\underline{F} \leq \lim_{i \to \infty} F_i \leq F_1 - \varepsilon_0 \varepsilon_1 \underline{c} \sum_{i=1}^{\infty} \parallel g_i \parallel^2$$

exists, so that

$$\sum_{i=1}^{\infty} \parallel g_i \parallel^2 \leq \frac{F_1 - \underline{F}}{\varepsilon_0 \varepsilon_1 \underline{c}}$$

and, necessarily, $\parallel g_i \parallel \to 0$.

Corollary Steepest descent method is globally convergent.

Remark 7 Assertion $\parallel g_i \parallel \to 0$ is stronger than the condition used in Definition 14. For proving global convergence in the sense of Definition 14, it suffices that (S1b) holds for an infinite subset of the set N.

Remark 8 If the basic optimization method is globally convergent (in the sense of Definition 14), then $x_i \to x^*$ does not necessarily hold. If the function $F : R^n \to R$ satisfies condition (F2), then the sequence $x_i \in R^n$, $i \in N$ does not diverge. However, it can have several cluster points. If some cluster point $x^* \in R^n$ of the sequence $x_i \in R^n$, $i \in N$, satisfies sufficient conditions for local minimum (Proposition 2), then $x_i \to x^*$ holds. In the subsequent considerations, we will suppose that global convergence automatically means that $x_i \to x^*$.

Theorem 6 (Linear convergence) Let $x_i \in R^n$, $i \in N$, be a sequence generated by an uniformly descent line search method, such that $x_i \to x^*$. Let the function $F : R^n \to R$ satisfy the conditions (F3) and (F4). Then for an arbitrary index $k \in N$ it holds

$$\frac{\parallel x_i - x^* \parallel}{\parallel x_k - x^* \parallel} \leq \sqrt{\frac{\overline{G}}{\underline{G}}}\; q^{i-k}$$

$\forall i \geq k$, where $q = \sqrt{1 - 2\varepsilon_0 \varepsilon_1 \underline{c} \underline{G}^2 / \overline{G}}$.

Proof (for (S3b)) Using (b), we can write

$$F_{i+1} - F^* \le F_i - F^* - \varepsilon_0 \varepsilon_1 \underline{c} \parallel g_i \parallel^2 .$$

From the mean value theorems we obtain

$$F_i - F^* \le \frac{1}{2} \overline{G} \parallel x_i - x^* \parallel^2, \tag{c}$$

$$F_i - F^* \ge \frac{1}{2} \underline{G} \parallel x_i - x^* \parallel^2, \tag{d}$$

$$\parallel g_i \parallel \le \overline{G} \parallel x_i - x^* \parallel, \tag{e}$$

$$\parallel g_i \parallel \ge \underline{G} \parallel x_i - x^* \parallel . \tag{f}$$

Therefore, we can write

$$\frac{F_{i+1} - F^*}{F_i - F^*} \le 1 - \varepsilon_0 \varepsilon_1 \underline{c} \frac{\parallel g_i \parallel^2}{F_i - F^*} \le 1 - \varepsilon_0 \varepsilon_1 \underline{c} \frac{2 \underline{G}^2 \parallel x_i - x^* \parallel^2}{\overline{G} \parallel x_i - x^* \parallel^2} = 1 - 2 \varepsilon_0 \varepsilon_1 \underline{c} \underline{G}^2 / \overline{G} = q^2 \qquad (g).$$

The condition (S3b) implies $\underline{c} = \varepsilon_0 (1 - \varepsilon_2) / \overline{G}$ (see proof of Theorem 5), so that

$$q^2 = 1 - 2 \varepsilon_0^2 \varepsilon_1 (1 - \varepsilon_2) \underline{G}^2 / \overline{G}^2 > 1 - 2 \varepsilon_0^2 \varepsilon_1 (1 - \varepsilon_1) \underline{G}^2 / \overline{G}^2 > 1 - \varepsilon_0^2 \underline{G}^2 / \overline{G}^2 \ge 0,$$

since $0 < \varepsilon_0 \le 1$, $0 < \varepsilon_1 < 1/2$, $\varepsilon_1 < \varepsilon_2 < 1$. Using the inequality (g) successively, we get

$$\frac{F_i - F^*}{F_k - F^*} \le q^{2(i-k)},$$

which together with (c) and (d) gives

$$\frac{\parallel x_i - x^* \parallel}{\parallel x_k - x^* \parallel} \le \sqrt{\frac{\overline{G}}{\underline{G}}} \sqrt{\frac{F_i - F^*}{F_k - F^*}} \le \sqrt{\frac{\overline{G}}{\underline{G}}} q^{i-k}.$$

Remark 9 The condition (F4) is usually supposed to hold from some index k (e.g. if $x_i \to x^*$ and x^* satisfies sufficient conditions for a local minimum, then we suppose that (F4) holds in a neighborhood of the point x^*).

Remark 10 Theorem 6 implies $\parallel e_{i+1} \parallel = O(\parallel e_i \parallel)$. This relation holds even if $\varepsilon_0 = 0$ (so that $q = 1$), or for an arbitrary descent line search method.

Definition 18 We say that the line search is asymptotically perfect if

$$\alpha_i = - \frac{s_i^T g_i}{s_i^T G^* s_i} (1 + O(\parallel e_i \parallel)).$$

Lemma 2 Let $x_i \in R^n$, $i \in N$, be a sequence generated by a uniformly descent line search method with asymptotically perfect line search, such that $x_i \to x^*$. Let the function $F : R^n \to R$ satisfy the conditions (F3)-(F5). Then

$$F_{i+1} - F_i = - \frac{1}{2} \frac{(s_i^T g_i)^2}{s_i^T G^* s_i} (1 + O(\parallel e_i \parallel)).$$

Proof We will write $u_i \sim v_i$ if $u_i = O(v_i)$ and $v_i = O(u_i)$. Using Theorem 2, we get

$$g_i = G^* e_i + O(\| e_i \|^2),$$

which together with (F3) and (F4) gives $g_i \sim e_i$. Since from (F3) and (F4) it follows that $s_i^T G^* s_i \sim \| s_i \|^2$, from (S1b) it follows that $s_i^T g_i \sim \| s_i \| \| g_i \|$ and since we suppose that $\| e_i \| \to 0$, we can write

$$\alpha_i = -\frac{s_i^T g_i}{s_i^T G^* s_i}(1 + O(\| e_i \|)) \sim \frac{\| g_i \|}{\| s_i \|},$$

so that $\| d_i \| = \| \alpha_i s_i \| \sim \| g_i \| \sim \| e_i \|$ and from Theorem 1

$$\begin{aligned} F_{i+1} - F_i &= \alpha_i s_i^T g_i + \frac{1}{2}\alpha_i^2 s_i^T G_i s_i + O(\| d_i \|^3) = \\ &= \alpha_i s_i^T g_i + \frac{1}{2}\alpha_i^2 s_i^T G^* s_i + \frac{1}{2} d_i^T (G_i - G^*) d_i + O(\| d_i \|^3) = \\ &= \alpha_i s_i^T g_i + \frac{1}{2}\alpha_i^2 s_i^T G^* s_i + O(\| e_i \|^3). \end{aligned}$$

Substituting the asymptotic perfect line search formula into this expression, we get

$$F_{i+1} - F_i = -\frac{1}{2}\frac{(s_i^T g_i)^2}{s_i^T G^* s_i}(1 + O(\| e_i \|)),$$

since $(s_i^T g_i)^2 / s_i^T G^* s_i \sim \| g_i \|^2 \sim \| e_i \|^2$.

Lemma 3 Let B be a symmetric positive definite (SPD) matrix. If the vectors $u \in R^n$ and $v \in R^n$ satisfy the condition

$$\frac{(u^T v)^2}{u^T u v^T v} \le \varepsilon^2,$$

where $0 \le \varepsilon \le 1$, then

$$\frac{(u^T B v)^2}{u^T B u v^T B v} \le \left(\frac{\kappa(B) - 1 + (\kappa(B) + 1)\varepsilon}{\kappa(B) + 1 + (\kappa(B) - 1)\varepsilon} \right)^2.$$

If the vectors $u \in R^n$ and $v \in R^n$ satisfy the condition

$$\frac{(u^T v)^2}{u^T u v^T v} \ge 1 - \varepsilon^2,$$

where $0 \le \varepsilon \le 1$, then

$$\frac{(u^T v)^2}{u^T B u v^T B^{-1} v} \ge \frac{4\kappa(B)(1 - \varepsilon^2)}{(\kappa(B) + 1 + (\kappa(B) - 1)\varepsilon)^2}$$

(here $\kappa(B)$ is the spectral condition number of the matrix B).

Proof (a) Let the vectors $u \in R^n$ and $v \in R^n$ satisfy the condition $(u^T v)^2/(u^T u v^T v) \le \varepsilon^2$. Without loss of generality, we assume that $\| u \| = 1$, $\| v \| = 1$ and will use the notation $V = [u, v]$. Let the vector w be a linear combination of the vectors u a v, so that $\| w \| = 1$ and $u^T w = 0$. Then numbers α and β exist, such that

$$v = \alpha u + \beta w$$

and, if we take into account that $\| u \| = 1$ and $\| w \| = 1$, we get $u^T v = \alpha$ and $v^T v = \alpha^2 + \beta^2$. The inequality $(u^T v)^2/(u^T u v^T v) \le \varepsilon^2$ and $\| v \| = 1$ imply

$$\alpha^2 \le \varepsilon^2$$

and

$$\alpha^2 + \beta^2 = 1$$

Set $W = [u, w]$. Then obviously $V = WM$ holds, where

$$M = \begin{bmatrix} 1, & \alpha \\ 0, & \beta \end{bmatrix}.$$

Since $V^T BV = M^T W^T BWM$, we can write

$$\kappa(V^T BV) \le \kappa(M^T M)\kappa(W^T BW).$$

Since the vector w was chosen in such a way that $W^T W = I$, we obtain

$$\frac{x^T W^T BWx}{x^T x} = \frac{x^T W^T BWx}{x^T W^T Wx} = \frac{y^T By}{y^T y},$$

where $y = Wx$, so that necessarily $\underline{\lambda}(W^T BW) = \underline{\lambda}(B)$, $\overline{\lambda}(W^T BW) = \overline{\lambda}(B)$ and

$$\kappa(W^T BW) = \frac{\overline{\lambda}(W^T BW)}{\underline{\lambda}(W^T BW)} = \frac{\overline{\lambda}(B)}{\underline{\lambda}(B)} = \kappa(B).$$

Since $\alpha^2 \le \varepsilon^2$ and $\alpha^2 + \beta^2 = 1$, it holds

$$M^T M = \begin{bmatrix} 1, & \alpha \\ \alpha, & 1 \end{bmatrix},$$

so that $\underline{\lambda}(M^T M) = 1 - |\alpha|$, $\overline{\lambda}(M^T M) = 1 + |\alpha|$ and

$$\kappa(M^T M) = \frac{\overline{\lambda}(M^T M)}{\underline{\lambda}(M^T M)} = \frac{1 + |\alpha|}{1 - |\alpha|} = \frac{1 + \varepsilon}{1 - \varepsilon}.$$

Therefore, we can write

$$\kappa(V^T BV) \le \kappa(M^T M)\kappa(W^T BW) = \kappa(B)\frac{1 + \varepsilon}{1 - \varepsilon}.$$

Let $\underline{\lambda}$ and $\overline{\lambda}$ be the eigenvalues of the matrix $V^T BV$ sorted by their sizes. Then

$$\det(V^T BV) = \underline{\lambda}\overline{\lambda} = \underline{\lambda}^2 \kappa(V^T BV).$$

Inequality $\left(\sqrt{u^T Bu} - \sqrt{v^T Bv}\right)^2 \ge 0$ implies that

$$\sqrt{u^T Buv^T Bv} \le \frac{1}{2}(u^T Bu + v^T Bv) = \frac{1}{2}Tr(V^T BV) = \frac{1}{2}(\underline{\lambda} + \overline{\lambda}) = \frac{1}{2}\underline{\lambda}(1 + \kappa(V^T BV)).$$

Therefore, we can write

$$\begin{aligned}
\frac{(u^T Bv)^2}{u^T Buv^T Bv} &= 1 - \frac{\det(V^T BV)}{u^T Buv^T Bv} \le 1 - \frac{4\kappa(V^T BV)}{(1 + \kappa(V^T BV))^2} = \\
&= \left(\frac{\kappa(V^T BV) - 1}{\kappa(V^T BV) + 1}\right)^2 \le \left(\frac{\kappa(B)\frac{1+\varepsilon}{1-\varepsilon} - 1}{\kappa(B)\frac{1+\varepsilon}{1-\varepsilon} + 1}\right)^2 = \\
&= \left(\frac{\kappa(B) - 1 + (\kappa(B) + 1)\varepsilon}{\kappa(B) + 1 + (\kappa(B) - 1)\varepsilon}\right)^2
\end{aligned}$$

(function $(x - 1)/(x + 1)$ is increasing for positive x).

(b) Let the vectors $u \in R^n$ and $v \in R^n$ satisfy the condition $(u^T v)^2/(u^T uv^T v) \ge 1 - \varepsilon^2$. Set $w = BHv$, where

$$H = B^{-1} - u(u^T Bu)^{-1} u^T.$$

Then

$$u^T w = u^T B(B^{-1} - u(u^T Bu)^{-1} u^T)v = u^T v - u^T Bu(u^T Bu)^{-1} u^T v = 0,$$

so that the vectors u and w are mutually orthogonal. Choose, in R^n, an orthogonal basis v_i, $1 \le i \le n$, such that $v_1 = u/ \parallel u \parallel$ and $v_2 = w/ \parallel w \parallel$ hold. Then

$$v = \sum_{i=1}^{n} (v^T v_i) v_i$$

and

$$v^T v = \sum_{i=1}^{n} (v^T v_i)^2 \ge (v^T v_1)^2 + (v^T v_2)^2 = \frac{(v^T u)^2}{u^T u} + \frac{(v^T w)^2}{w^T w},$$

so that

$$\frac{(v^T w)^2}{w^T w v^T v} = 1 - \frac{(v^T u)^2}{u^T u v^T v} \le \varepsilon^2$$

and if we use (a), we get

$$\frac{(w^T B^{-1} v)^2}{w^T B^{-1} w v^T B^{-1} v} \le \left(\frac{\kappa(B) - 1 + (\kappa(B) + 1)\varepsilon}{\kappa(B) + 1 + (\kappa(B) - 1)\varepsilon} \right)^2.$$

On the other hand (according to the definition of the matrix H and the vector w and from the orthogonality $u^T w = 0$) it holds

$$\begin{aligned} w^T B^{-1} w &= w^T B^{-1} B H v = w^T B^{-1} v - w^T u (u^T Bu)^{-1} u^T v \\ &= w^T B^{-1} v = v^T H B B^{-1} v = v^T H v \end{aligned}$$

and

$$v^T H v = v^T B^{-1} v - (u^T v)^2 (u^T Bu)^{-1},$$

so that

$$\begin{aligned} \frac{(u^T v)^2}{u^T B u v^T B^{-1} v} &= 1 - \frac{v^T H v}{v^T B^{-1} v} = 1 - \frac{(w^T B^{-1} v)^2}{w^T B^{-1} w v^T B^{-1} v} \ge \\ &\ge 1 - \left(\frac{\kappa(B) - 1 + (\kappa(B) + 1)\varepsilon}{\kappa(B) + 1 + (\kappa(B) - 1)\varepsilon} \right)^2 = \frac{4\kappa(B)(1 - \varepsilon^2}{(\kappa(B) + 1 + (\kappa(B) - 1)\varepsilon)^2)}. \end{aligned}$$

Theorem 7 Let the assumptions of Lemma 2 be satisfied. Then

$$\limsup_{i \to \infty} \parallel x_i - x^* \parallel^{\frac{1}{i}} \le \frac{\kappa(G^*) - 1 + (\kappa(G^*) + 1)\sqrt{1 - \varepsilon_0^2}}{\kappa(G^*) + 1 + (\kappa(G^*) - 1)\sqrt{1 - \varepsilon_0^2}}$$

.

Proof Theorem 2 implies

$$F_i - F^* = \frac{1}{2} e_i^T G^* e_i + O(\parallel e_i \parallel^3),$$

$$g_i = G^* e_i + O(\parallel e_i \parallel^2),$$

so that using (F3) and (F4) and the fact that $\parallel g_i \parallel \sim \parallel e_i \parallel$ we get

$$e_i = (G^*)^{-1} g_i (1 + O(\| e_i \|))$$

and

$$F_i - F^* = \frac{1}{2} g_i^T (G^*)^{-1} g_i (1 + O(\| e_i \|)),$$

Applying Lemma 2, we can write

$$\frac{F_{i+1} - F^*}{F_i - F^*} = 1 + \frac{F_{i+1} - F_i}{F_i - F^*} = 1 - \frac{(s_i^T g_i)^2}{s_i^T G^* s_i g_i^T (G^*)^{-1} g_i} (1 + O(\| e_i \|))$$

(since $(1 + O(\| e_i \|))/(1 + O(\| e_i \|)) = 1 + O(\| e_i \|)$ holds). Using (S1b), we get $(s_i^T g_i)^2 \geq \varepsilon_0^2 \| s_i \|^2 \| g_i \|^2$, so that using Lemma 3 we can write

$$\frac{(s_i^T g_i)^2}{s_i^T G^* s_i g_i^T (G^*)^{-1} g_i} \geq \frac{4\kappa(G^*)\varepsilon_0^2}{(\kappa(G^*) + 1 + (\kappa(G^*) - 1)\sqrt{1 - \varepsilon_0^2})^2},$$

so that

$$\begin{aligned} \frac{F_{i+1} - F^*}{F_i - F^*} &\leq \left(\frac{(\kappa(G^*) - 1 + (\kappa(G^*) + 1)\sqrt{1 - \varepsilon_0^2}}{(\kappa(G^*) + 1 + (\kappa(G^*) - 1)\sqrt{1 - \varepsilon_0^2}} \right)^2 (1 + O(\| e_i \|) = \\ &= \hat{q}^2 (1 + O(\| e_i \|)). \end{aligned}$$

Therefore, for an arbitrary number $q, \hat{q} < q < 1$, an index $k \in N$ exists, so that

$$\frac{F_{i+1} - F^*}{F_i - F^*} \leq q^2$$

$\forall i \geq k$. Hence we can proceed by the same way as in proof of Theorem 6, so that

$$\frac{F_i - F^*}{F_k - F^*} \leq q^{2(i-k)}$$

and

$$\frac{\| x_i - x^* \|}{\| x_k - x^* \|} \leq \sqrt{\frac{\overline{G}}{\underline{G}}} q^{i-k}$$

and using Theorem 2 we obtain

$$\limsup_{i \to \infty} \| x_i - x^* \|^{\frac{1}{i}} \leq q.$$

Since the above holds for an arbitrary number q, $\hat{q} < q < 1$, the theorem is proved.

Remark 11 If the steepest descent method is used, then $\varepsilon_0 = 1$, so that

$$\limsup_{i \to \infty} \leq \frac{\kappa(G^*) - 1}{\kappa(G^*) + 1}.$$

Remark 12 An asymptotically perfect line search is realized if we choose a stepsize using quadratic or cubic interpolation (see Remark 15).

Theorem 8 (Superlinear convergence). Let the function $F : R^n \to R$ satisfy the conditions (F3) and (F4). Let $x_i \in R^n$, $i \in N$, be a sequence generated by the line search method, such that $x_i \to x^*$, and in addition let

$$\lim_{i \to \infty} \frac{\| B_i s_i + g_i \|}{\| g_i \|} = 0 \qquad (\alpha)$$

and

$$\lim_{i\to\infty} \frac{\| (B_i - G_i)s_i \|}{\| s_i \|} = 0. \qquad (\beta)$$

Let $\alpha_i = 1$, whenever this value satisfy the conditions (S2) and (S3). Then the condition (S1b) holds (line search is uniformly descent) and an index $k \in N$ exists such that $\alpha_i = 1$, $\forall i \geq k$, and the sequence x_i, $i \in N$, converges superlinearly to the point $x^* \in R^n$.

Proof (a) We show that index $k_1 \in N$ exists, so that

$$\| g_i \| / \overline{G} \leq \| s_i \| \leq \| g_i \| / \underline{G}$$

$\forall i \geq k_1$, if $\underline{G} < \underline{\lambda}(G^*)$ and $\overline{G} > \overline{\lambda}(G^*)$. Denote $\omega_i = (B_i s_i + g_i)/ \| g_i \|$ and $\vartheta_i = (B_i - G_i)s_i / \| s_i \|$. then

$$G_i s_i = (B_i s_i + g_i) - (B_i - G_i)s_i - g_i = \omega_i \| g_i \| - \vartheta_i \| s_i \| - g_i,$$

so that

$$\| s_i \| \geq \frac{1 - \| \omega_i \|}{\| G_i \| + \| \vartheta_i \|} \| g_i \|$$

and, since $\| \vartheta_i \| \to 0$ and $\| \omega_i \| \to 0$ (by (α) and (β)) and $\| G_i \| \to \| G^* \| = \overline{\lambda}(G^*) < \overline{G}$, an index $k_0 \in N$ exists so that $\| s_i \| \geq \| g_i \| / \overline{G}$ $\forall i \geq k_0$. Similarly

$$s_i = G_i^{-1}(\omega_i \| g_i \| - \vartheta_i \| s_i \| - g_i),$$

so that

$$\| s_i \| \leq \frac{\| G_i^{-1} \| (1 + \| \omega_i \|)}{1 - \| G_i^{-1} \| \| \vartheta_i \|} \| g_i \|$$

and, since $\| \vartheta_i \| \to 0$ and $\| \omega_i \| \to 0$ (by (α) and (β)) and $\| G_i^{-1} \| \to \| (G^*)^{-1} \| = 1/\underline{\lambda}(G^*) < 1/\underline{G}$, an index $k_1 \geq k_0$ exists, so that $\| s_i \| \leq \| g_i \| / \underline{G}$ $\forall i \geq k_1$.
(b) We show that an index $k_2 \geq k_1$ exists, such that $-s_i^T g_i \geq \tilde{\varepsilon}_0 \| s_i \| \| g_i \|$ $\forall i \geq k_2$, if $\tilde{\varepsilon}_0 < 1/\kappa(G^*)$. Choose $\underline{G} < \underline{\lambda}(G^*)$ and $\overline{G} > \overline{\lambda}(G^*)$ so that $\tilde{\varepsilon}_0 = \underline{G}/\overline{G}$ holds. The definition of ω_i and ϑ_i and (a) imply that

$$\begin{aligned} -s_i^T g_i &= s_i^T(G_i s_i + (B_i - G_i)s_i - (B_i s_i + g_i)) \geq (\underline{\lambda}(G_i) - \| \vartheta_i \|) \| s_i \|^2 - \| \omega_i \| \| s_i \| \| g_i \| \geq \\ &\geq (\underline{\lambda}(G_i)/\overline{G} - \| \vartheta_i \| / \underline{G} - \| \omega_i \|) \| s_i \| \| g_i \| \end{aligned}$$

and, since $\| \omega_i \| \to 0$ and $\| \vartheta_i \| \to 0$ (by (α) and (β)) and $\underline{\lambda}(G_i) \to \underline{\lambda}(G^*) > \underline{G}$, an index $k_2 \geq k_1$ exists, so that $-s_i^T g_i \geq (\underline{G}/\overline{G}) \| s_i \| \| g_i \| = \tilde{\varepsilon}_0 \| s_i \| \| g_i \|$ $\forall i \geq k_2$. If we set

$$\varepsilon_0 = \min(\tilde{\varepsilon}_0, \min_{1 \leq i < k_2} (-s_i^T g_i / \| s_i \| \| g_i \|)),$$

then (S1b) holds.
(c) We show that an index $k \geq k_2$ exists, so that the value $\alpha_i = 1$ satisfies the conditions (S2) and (S3b). Denote

$$\eta_i = \frac{s_i^T g_i + s_i^T G_i s_i}{s_i^T g_i}.$$

Then, using previous results, we get for $i \geq k_2$

$$|\eta_i| = \frac{|s_i^T g_i + s_i^T G_i s_i|}{|s_i^T g_i|} \leq \frac{\| s_i \| \| g_i + G_i s_i \|}{\varepsilon_0 \| s_i \| \| g_i \|} \leq \frac{\overline{G}}{\underline{G}} \left(\frac{\| g_i + B_i s_i \|}{\| g_i \|} + \frac{\| (B_i - G_i)s_i \|}{\| g_i \|} \right)$$

and (α), (β) and (a) imply $|\eta_i| \to 0$. Now we use the mean value theorems

$$F(x_i+s_i)-F(x_i)=s_i^Tg_i+\frac{1}{2}s_i^TG_is_i+o(\| s_i \|^2),$$

$$s_i^Tg(x_i+s_i)=s_i^Tg_i+s_i^TG_is_i+o(\| s_i \|^2).$$

Therefore, we can write

$$\lim_{i\to\infty}\frac{F(x_i+s_i)-F(x_i)}{s_i^Tg_i}=\frac{1}{2}+\lim_{i\to\infty}(\frac{1}{2}\eta_i+o(\| s_i \|^2)/\| s_i \|^2)=\frac{1}{2},$$

$$\lim_{i\to\infty}\frac{s_i^Tg(x_i+s_i)}{s_i^Tg_i}=\lim_{i\to\infty}(\eta_i+o(\| s_i \|^2)/\| s_i \|^2)=0,$$

and, since $0<\varepsilon_1<1/2$ and $\varepsilon_1<\varepsilon_2<1$, an index $k\geq k_2$ exists, so that (S2) and (S3b) (with $\alpha_i=1$) hold $\forall i\geq k$.
(d) Superlinear convergence. We use the mean value theorem

$$g(x_i+s_i)=g_i+G_is_i+o(\| s_i \|).$$

Using the previous results, we get $x_{i+1}=x_i+s_i\ \forall i\geq k$ and $\| s_i \|\sim\| g_i \|\to 0$.
Therefore, we can write

$$\begin{aligned}\frac{\| x_{i+1}-x^* \|}{\| x_i-x^* \|} &\leq \frac{\overline{G}}{\underline{G}}\frac{\| g_{i+1} \|}{\| g_i \|}\leq\\ &\leq \frac{\overline{G}}{\underline{G}}\left(\frac{\| g(x_i+s_i)-g_i-B_is_i \|}{\| g_i \|}+\frac{\| B_is_i+g_i \|}{\| g_i \|}\right)\leq\\ &\leq \frac{\overline{G}}{\underline{G}}\left(\frac{\| (B_i-G_i)s_i \|}{\| g_i \|}+\frac{\| B_is_i+g_i \|}{\| g_i \|}+o(\| s_i \|)/\| s_i \|\right),\end{aligned}$$

so that (α), (β) and (a) imply

$$\lim_{i\to\infty}\frac{\| x_{i+1}-x^* \|}{\| x_i-x^* \|}=0$$

Now we focus our attention to the implementation of the line search methods. First we describe an algorithm for the stepsize selection.

Algorithm 1 (S3b) Data $0<\beta_1<\beta_2<1<\gamma_1<\gamma_2$.

Step 1 Choose an initial stepsize $\alpha>0$. Set $\overline{\alpha}=0$.

Step 2 Set $\underline{\alpha}=\overline{\alpha}$ and $\overline{\alpha}=\alpha$. If the conditions (S2) and (S3b) are simultaneously satisfied, then terminate the computation with $\alpha_i=\alpha$. If the condition (S2) does not hold, then go to step 4.

Step 3 Determine the value α, using an extrapolation, in such a way that $\gamma_1\overline{\alpha}\leq\alpha\leq\gamma_2\overline{\alpha}$ holds. Go to step 2.

Step 4 Determine the value α, using an interpolation, in such a way that $\beta_1(\overline{\alpha}-\underline{\alpha})\leq(\alpha-\underline{\alpha})\leq\beta_2(\overline{\alpha}-\underline{\alpha})$ holds.

Step 5 If the conditions (S2) and (S3b) are simultaneously satisfied, then terminate the computation with $\alpha_i=\alpha$. If the condition (S2) does not hold, set $\overline{\alpha}=\alpha$, otherwise set $\underline{\alpha}=\alpha$. Go to step 4.

Remark 13 If the conditions (F1) and (F3) are satisfied, then Algorithm 1 finds a stepsize satisfying the conditions (S2) and (S3b) after a finite number of steps. If, in addition, the condition (F4) holds, then the number of steps does not depend on the index $i\in N$.

Remark 14 Extrapolation and interpolation. Denote $\varphi(\alpha) = F(x_i + \alpha s_i)$, $\varphi'(\alpha) = s_i^T g(x_i + \alpha s_i)$ and

$$A = \frac{\varphi(\overline{\alpha}) - \varphi(\underline{\alpha})}{(\overline{\alpha} - \underline{\alpha})\varphi'(\underline{\alpha})},$$

$$B = \frac{\varphi'(\overline{\alpha})}{\varphi'(\underline{\alpha})}.$$

Quadratic interpolation (two values):

$$\alpha - \underline{\alpha} = \frac{\overline{\alpha} - \underline{\alpha}}{2(1 - A)}. \tag{Ka}$$

Quadratic interpolation (two derivatives):

$$\alpha - \underline{\alpha} = \frac{\overline{\alpha} - \underline{\alpha}}{1 - B}. \tag{Kb}$$

Cubic interpolation :

$$\alpha - \underline{\alpha} = \frac{\overline{\alpha} - \underline{\alpha}}{D + \sqrt{D^2 - 3C}}, \tag{C}$$

where

$$C = (B - 1) - 2(A - 1),$$

$$D = (B - 1) - 3(A - 1).$$

Remark 15 If the stepsize is determined by (Ka) or (Kb) or (C) and (F3) and (F4) hold, then the line search is asymptoytically perfect.

Remark 16 Initial stepsize selection. For superlinearly convergent methods we choose $\alpha = 1$. For conjugate gradient methods we choose $\alpha = 2(F_{i-1} - F_i)/s_i^T g_i$ or (in the first iteration) $\alpha = 2(\underline{F} - F_i)/s_i^T g_i$.

Summary: For uniformly descent line search methods (with the parameters $\varepsilon_0, \varepsilon_1, \varepsilon_2$ in the conditions (S1), (S2), (S3)) the following implications hold:

a) (F1) - (F3) $\Rightarrow$ global convergence.
b) (F3) - (F4) $\Rightarrow$ linear convergence.
c) (F3) - (F5) $\Rightarrow$ asymptotic rate of convergence

$$\limsup_{i \to \infty} \| x_i - x^* \|^{\frac{1}{i}} \le \frac{\kappa(G^*) - 1 + (\kappa(G^*) + 1)\sqrt{1 - \varepsilon_0^2}}{\kappa(G^*) + 1 + (\kappa(G^*) - 1)\sqrt{1 - \varepsilon_0^2}}.$$

d) (F3) - (F4) $\Rightarrow$ superlinear convergence when

$$\frac{\| B_i s_i + g_i \|}{\| g_i \|} \to 0,$$

$$\frac{\| (G^* - B_i) s_i \|}{\| g_i \|} \to 0.$$

2.2 Conjugate gradient methods

Definition 19 We say that a basic optimization method is a conjugate gradient method if $s_1 = -g_1$ and

$$s_{i+1} = -g_{i+1} + \frac{y_i^T g_{i+1}}{y_i^T s_i} s_i \tag{CGa}$$

(Sorenson) or

$$s_{i+1} = -g_{i+1} + \frac{y_i^T g_{i+1}}{g_i^T g_i} s_i \tag{CGb}$$

(Polak-Ribiere) or

$$s_{i+1} = -g_{i+1} + \frac{g_{i+1}^T g_{i+1}}{g_i^T g_i} s_i \tag{CGc}$$

(Fletcher-Reeves) for $i \in N$. Here $y_i = g_{i+1} - g_i$.

Remark 17 Method (CGa) is theoretically the most sophisticated one. Method (CGb) gives the best practical results. Method (CGc) is the most simple one and it is globally convergent without restarts.

Theorem 9 (Global convergence). Let the function $F : R^n \to R$ satisfy the condition (F1)-(F3). Then the Fletcher-Reeves conjugate gradient method (CGc) with a stepsize selection satisfying the strong Wolfe condition (S2) and (S3a), where $\varepsilon_1 < \varepsilon_2 < 1/2$, is globally convergent.

Proof (a) (Al-Baali) We prove by induction the inequality

$$-1 - \frac{\varepsilon_2}{1-\varepsilon_2} \le \frac{s_i^T g_i}{\| g_i \|^2} \le -1 + \frac{\varepsilon_2}{1-\varepsilon_2}$$

$\forall i \in N$. For $i = 1$ the inequality is true, since $s_1 = -g_1$ and, therefore, $s_1^T g_1 / \| g_1 \|^2 = -1$. If we use (CGc), we obtain

$$\frac{s_{i+1}^T g_{i+1}}{\| g_{i+1} \|^2} = -1 + \frac{\| g_{i+1} \|^2}{\| g_i \|^2} \frac{s_i^T g_{i+1}}{\| g_{i+1} \|^2} = -1 + \frac{s_i^T g_{i+1}}{\| g_i \|^2}.$$

By (S3a) it holds

$$|s_i^T g_{i+1}| \le -\varepsilon_2 s_i^T g_i,$$

so that

$$\begin{aligned}
-1 + \varepsilon_2 \frac{s_i^T g_i}{\| g_i \|^2} &\le \frac{s_{i+1}^T g_{i+1}}{\| g_{i+1} \|^2} \le -1 - \varepsilon_2 \frac{s_i^T g_i}{\| g_i \|^2} \\
-1 - \varepsilon_2 \left(1 + \frac{\varepsilon_2}{1-\varepsilon_2}\right) &\le \frac{s_{i+1}^T g_{i+1}}{\| g_{i+1} \|^2} \le -1 + \varepsilon_2 \left(1 + \frac{\varepsilon_2}{1-\varepsilon_2}\right) \\
-1 - \frac{\varepsilon_2}{1-\varepsilon_2} &\le \frac{s_{i+1}^T g_{i+1}}{\| g_{i+1} \|^2} \le -1 + \frac{\varepsilon_2}{1-\varepsilon_2}.
\end{aligned}$$

(b) From (S3a) it follows the truth of (S3b). Using (F3) and (S3b), we get

$$\varepsilon_2 s_i^T g_i \le s_i^T g_{i+1} \le s_i^T g_i + \alpha_i \overline{G} \| s_i \|^2,$$

so that

$$\alpha_i \ge -\frac{(1-\varepsilon_2) s_i^T g_i}{\overline{G} \| s_i \|^2}.$$

This together with (S2) implies

$$F_i - F_{i+1} \geq -\varepsilon_1 \alpha_i s_i^T g_i \geq \frac{\varepsilon_1(1-\varepsilon_2)(s_i^T g_i)^2}{\overline{G} \parallel s_i \parallel^2} = \frac{\varepsilon_1(1-\varepsilon_2)}{\overline{G}} \frac{(s_i^T g_i)^2}{\parallel g_i \parallel^4} \frac{\parallel g_i \parallel^4}{\parallel s_i \parallel^2}.$$

From the Al-Baali inequality (right hand side) it follows

$$-\frac{s_i^T g_i}{\parallel g_i \parallel^2} \geq 1 - \frac{\varepsilon_2}{1-\varepsilon_2} = \frac{1-2\varepsilon_2}{1-\varepsilon_2} > 0,$$

since $\varepsilon_1 < \varepsilon_2 < 1/2$. Therefore

$$F_1 - \underline{F} \geq \lim_{i\to\infty}(F_1 - F_{i+1}) = \sum_{i=1}^{\infty}(F_i - F_{i+1}) \geq \sum_{i=1}^{\infty} \frac{\varepsilon_1(1-2\varepsilon_2)^2}{\overline{G}(1-\varepsilon_2)} \frac{\parallel g_i \parallel^4}{\parallel s_i \parallel^2}.$$

Suppose that

$$\liminf_{i\to\infty} \parallel g_i \parallel = 0$$

does not hold. Then $\underline{\varepsilon} > 0$ exists, so that $\parallel g_i \parallel \geq \underline{\varepsilon}\ \forall i \in N$, so that

$$\infty > \frac{(F_1 - \underline{F})\overline{G}(1-\varepsilon_2)}{\underline{\varepsilon}^4 \varepsilon_1 (1-2\varepsilon_2)^2} \geq \sum_{i=1}^{\infty} \frac{1}{\parallel s_i \parallel^2}.$$

(c) From (S3a) follows (S3b). Using (S3b) and the Al-Baali inequality (left hand side), we get

$$-s_i^T g_{i+1} \leq -\varepsilon_2 s_i^T g_i \leq \varepsilon_2 \left(1 + \frac{\varepsilon_2}{1-\varepsilon_2}\right) \parallel g_i \parallel^2 = \frac{\varepsilon_2}{1-\varepsilon_2} \parallel g_i \parallel^2 .$$

If we use (CGc), we can write

$$\begin{aligned}
\parallel s_{i+1} \parallel^2 &\leq \parallel g_{i+1} \parallel^2 - 2\frac{\parallel g_{i+1} \parallel^2}{\parallel g_i \parallel^2} s_i^T g_{i+1} + \frac{\parallel g_{i+1} \parallel^4}{\parallel g_i \parallel^4} \parallel s_i \parallel^2 \leq \\
&\leq \parallel g_{i+1} \parallel^2 + \frac{2\varepsilon_2}{1-\varepsilon_2} \parallel g_{i+1} \parallel^2 + \frac{\parallel g_{i+1} \parallel^4}{\parallel g_i \parallel^4} \parallel s_i \parallel^2 = \\
&= \frac{1+\varepsilon_2}{1-\varepsilon_2} \parallel g_{i+1} \parallel^2 + \frac{\parallel g_{i+1} \parallel^4}{\parallel g_i \parallel^4} \parallel s_i \parallel^2 .
\end{aligned}$$

This is a recurrence relation for $\parallel s_i \parallel$. By successive substitution, we get

$$\begin{aligned}
\parallel s_{i+1} \parallel^2 &\leq \frac{1+\varepsilon_2}{1-\varepsilon_2} \parallel g_{i+1} \parallel^2 + \frac{\parallel g_{i+1} \parallel^4}{\parallel g_i \parallel^4} \parallel s_i \parallel^2 \leq \\
&\leq \frac{1+\varepsilon_2}{1-\varepsilon_2} \parallel g_{i+1} \parallel^2 + \frac{\parallel g_{i+1} \parallel^4}{\parallel g_i \parallel^4} \left(\frac{1+\varepsilon_2}{1-\varepsilon_2} \parallel g_i \parallel^2 + \frac{\parallel g_i \parallel^4}{\parallel g_{i-1} \parallel^4} \parallel s_{i-1} \parallel^2\right) \leq \\
&\leq \frac{1+\varepsilon_2}{1-\varepsilon_2} \parallel g_{i+1} \parallel^4 \sum_{j=1}^{i+1} \frac{1}{\parallel g_j \parallel^2}.
\end{aligned}$$

Suppose that

$$\liminf_{i\to\infty} \parallel g_i \parallel = 0$$

does not hold. Then $0 < \underline{\varepsilon} < \overline{\varepsilon} < \infty$ exists, so that $\underline{\varepsilon} \leq \parallel g_i \parallel \leq \overline{\varepsilon}\ \forall i \in N$ (existence of $\overline{\varepsilon}$ follows from (F2) and (F3)), so that

$$\parallel s_i \parallel^2 \leq \frac{(1+\varepsilon_2)\overline{\varepsilon}^4}{(1-\varepsilon_2)\underline{\varepsilon}^2} i$$

and we can write

$$\sum_{i=1}^{\infty} \frac{1}{\| s_i \|^2} \ge \frac{(1-\varepsilon_2)\underline{\varepsilon}^2}{(1+\varepsilon_2)\overline{\varepsilon}^4} \sum_{i=1}^{\infty} \frac{1}{i} = \infty,$$

which is a contradiction, since in (b) we have proved that this sum is finite.

Remark 18 Denote β_i the coefficient in (CGa)-(CGc). It can be proved that the conjugate gradient method is globally convergent if

$$0 \le \beta_i \le \frac{1}{2\varepsilon_3} \frac{\| g_{i+1} \|^2}{\| g_i \|^2} \qquad \forall i \in N$$

where $0 < \varepsilon_2 < \varepsilon_3 < 1/2$. However, this cannot be guaranteed in general. Therefore, restart $(s_i = -g_i)$ is used if the above condition does not hold.

Theorem 10 (Quadratic termination) Let $Q : R^n \to R$ be a strictly convex quadratic function, $Q(x) = \frac{1}{2}(x-x^*)^T G(x-x^*)$. Let x_i, $i \in N$, be a sequence generated by a conjugate gradient method with a perfect line search ($s_i^T g_{i+1} = 0$ holds $\forall i \in N$). Then an index $k \le n$ exists, so that $g_{k+1} = 0$ and $x_{k+1} = x^*$.

Proof (For CGa). Assume that $g_i \ne 0$ $\forall 1 \le i \le n$. We prove by induction that $s_i \ne 0$ and $\alpha_i \ne 0$ $\forall 1 \le i \le n$ and that it holds

$$(\alpha) \qquad s_j^T g_i = 0 \qquad \forall 1 \le j < i \le n+1,$$

$$(\beta) \qquad g_j^T g_i = 0 \qquad \forall 1 \le j < i \le n+1,$$

$$(\gamma) \qquad s_j^T G s_i = 0 \qquad \forall 1 \le j < i \le n.$$

From (β) it follows that the gradients g_i, $1 \le i \le n$, are nonzero and mutually conjugate, hence linearly independent, so that necessarily $g_{n+1} = 0$.

For $i = 1$, $s_1^T g_1 = -g_1^T g_1 < 0$ holds, so that $s_1 \ne 0$ and $\alpha_1 \ne 0$ and nothing else has to be proved. The induction steps:

(a) Let $i \le n$. The inductive assumption (γ) and Q imply

$$s_j^T g_{i+1} = s_j^T g_i + s_j^T y_i = s_j^T g_i + \alpha_i s_j^T G s_i = 0$$

$\forall 1 \le j < i$, since $y_i = g_{i+1} - g_i = \alpha_i G s_i$ holds for the quadratic function $Q(x)$. Moreover, the perfect line search guarantee that $s_i^T g_{i+1} = 0$. Therefore, $s_j^T g_{i+1} = 0$ $\forall 1 \le j \le i$.

(b) Let $i \le n$. Using (CGa), we obtain

$$\begin{aligned} g_1 &= -s_1 \\ g_j &= -s_j + \beta_{j-1} s_{j-1} \qquad \forall 1 < j \le i, \end{aligned}$$

so that (a) implies

$$\begin{aligned} g_1^T g_{i+1} &= -s_1^T g_{i+1} = 0 \\ g_j^T g_{i+1} &= -s_j^T g_{i+1} + \beta_{j-1} s_{j-1}^T g_{i+1} = 0 \qquad \forall 1 < j \le i. \end{aligned}$$

(c) Let $i < n$. Using (CGa) and (a), we can write

$$s_{i+1}^T g_{i+1} = -g_{i+1}^T g_{i+1} + \frac{y_i^T g_{i+1}}{y_i^T s_i} s_i^T g_{i+1} = -g_{i+1}^T g_{i+1} < 0,$$

so that $s_{i+1} \neq 0$ and $\alpha_{i+1} \neq 0$. Using (CGa) and (b) we obtain

$$s_j^T G s_{i+1} = -s_j^T G g_{i+1} + \beta_i s_j^T G s_i = -s_j^T G g_{i+1} = -\frac{1}{\alpha_j} y_j^T g_{i+1} = -\frac{1}{\alpha_j}(g_{j+1} - g_j)^T g_{i+1} = 0$$

$\forall 1 \leq j < i$ since $s_j^T G s_i = 0$ $\forall 1 \leq j < i$ by the inductive assumption (γ). Moreover, (CGa) implies

$$s_i^T G s_{i+1} = -s_i^T G g_{i+1} + \frac{y_i^T g_{i+1}}{y_i^T s_i} s_i^T G s_i = -s_i^T G g_{i+1} + \frac{s_i^T G g_{i+1}}{y_i^T s_i} y_i^T s_i = 0,$$

so that $s_j^T G s_{i+1} = 0$ $\forall 1 \leq j \leq i$.

Remark 19 Proof was done for CGa. Theorem 10 holds also for other conjugate gradient methods since (β) implies

$$y_i^T g_{i+1} = g_{i+1}^T g_{i+1} - g_i^T g_{i+1} = g_{i+1}^T g_{i+1}$$

and from (α) it follows that

$$y_i^T s_i = g_{i+1}^T s_i - g_i^T s_i = -g_i^T s_i = g_i^T g_i - \beta_{i-1} g_i^T s_{i-1} = g_i^T g_i.$$

Remark 20 A conjugate gradient method with perfect line search finds a minimum of a quadratic function after at most n steps. However, this is not true if

- line search is not perfect,
- function is not quadratic,
- Hessian matrix is ill-conditioned and round-off errors destroy this property.

In such a case, we have to continue in the computation. But restart ($s_{n+1} = -g_{n+1}$) is necessary for assumptions of Theorem 10 to be satisfied.

Definition 20 We say that a basic optimization method is a restarted conjugate gradient method if $s_i = -g_i$ for $i \in M$ and one of the formulas (CGa), (CGb), (CGc) is used for $i \notin M$, where $M = \{nk + 1 : k \in N\}$.

Remark 21 The restarted conjugate gradient method is globally convergent (it suffices to modify the proof of Theorem 5 in the way noticed in remark 7).

Theorem 11 Let $x_i \in R^n$, $i \in N$, be a sequence generated by the restarted conjugate gradient method (CGc) with the stepsize selection satisfying the strong Wolfe condition (S2) and (S3a), where $\varepsilon_1 < \varepsilon_2 < 1/2$, such that $x_i \to x^*$. Let the function $F : R^n \to R$ satisfy the conditions (F3) and (F4). Then the direction vectors $\| s_i \|$, $i \in N$ are uniformly descent and $\| s_i \| \sim \| g_i \|$ holds.

Proof (a) Obviously $\| e_i \| = O(\| e_{i-1} \|)$ (Remark 10) and $\| g_i \| \sim \| e_i \|$ (Remark 2), so that $\| g_i \| = O(\| g_{i-1} \|)$. Therefore, a constant $c < \infty$ exists such that

$$\frac{\| g_i \|}{\| g_{i-1} \|} \leq c \qquad \forall i \notin M.$$

Let $i \notin M$. Then (CGc) implies

$$\| s_i \| \leq \| g_i \| + \frac{\| g_i \|^2}{\| g_{i-1} \|^2} \| s_{i-1} \|,$$

so that

$$\frac{\| s_i \|}{\| g_i \|} \leq 1 + \frac{\| g_i \|}{\| g_{i-1} \|} \frac{\| s_{i-1} \|}{\| g_{i-1} \|} \leq 1 + c \frac{\| s_{i-1} \|}{\| g_{i-1} \|}.$$

Let $k = \sup\{j \in M, j \le i\}$. Since $s_k = -g_k$, $\| s_k \| / \| g_k \| = 1$ holds, then, applying the last inequality successively, we obtain

$$\frac{\| s_i \|}{\| g_i \|} \le \sum_{j=0}^{i-k} c^j \le \sum_{j=0}^{n} c^j \triangleq \overline{c}.$$

(b) Using the Al-Baali inequality (left hand side), we get

$$-\frac{s_i^T g_i}{\| g_i \|^2} \ge 1 - \frac{\varepsilon_2}{1-\varepsilon_2} = \frac{1-2\varepsilon_2}{1-\varepsilon_2},$$

which together with (a) gives

$$-\frac{s_i^T g_i}{\| s_i \| \| g_i \|} = -\frac{s_i^T g_i}{\| g_i \|^2} \frac{\| g_i \|}{\| s_i \|} \ge -\frac{1}{\overline{c}} \frac{s_i^T g_i}{\| g_i \|^2} \ge \frac{1}{\overline{c}} \frac{1-2\varepsilon_2}{1-\varepsilon_2},$$

so that $-s_i^T g_i \ge \varepsilon_0 \| s_i \| \| g_i \|$, where $\varepsilon_0 = (1-2\varepsilon_2)/(\overline{c}(1-\varepsilon_2))$.
(c) Using the Al-Baali inequality together with the Schwartz inequality, we get

$$\| s_i \| \| g_i \| \ge -s_i^T g_i \ge \frac{1-2\varepsilon_2}{1-\varepsilon_2} \| g_i \|^2,$$

which gives $\| s_i \| \ge \underline{c} \| g_i \|$, kde $\underline{c} = (1-2\varepsilon_2)/(1-\varepsilon_2)$.

Proposition 3 (n-step quadratic convergence) Let the assumptions of Theorem 11 be satisfied and let the line search be asymptotically perfect. Let the function $F : R^n \to R$ satisfy in addition the condition (F5). Then an index $k \in M$ and a constant $\overline{C} < \infty$ exist, so that

$$\| x_{i+n} - x^* \| \le \overline{C} \| x_i - x^* \|^2$$

holds $\forall i \in M$, $i \ge k$.

Lemma 4 Let the assumptions of Theorem 10 be satisfied. Let $g_i \ne 0$ for some index $1 \le i \le n$. Then

$$\frac{Q(x_{i+1}) - Q(x^*)}{Q(x_1) - Q(x^*)} \le \max_{1 \le k \le n} \left(\overline{P}_i(\lambda_k)\right)^2$$

holds, where $\overline{P}_i(\lambda)$ is an arbitrary polynomial of degree i, such that $\overline{P}_i(0) = 1$, and λ_k, $1 \le k \le n$, are eigenvalues of the matrix G.

Proof (a) We prove by induction that for $1 \le j \le i$, $g_j \in \mathcal{K}_j$ and $s_j \in \mathcal{K}_j$ holds, where

$$\mathcal{K}_j = \mathrm{span}\{g_1, Gg_1, \ldots, G^{j-1} g_1\}$$

is the Krylov subspace of degree j generated by the matrix G and the vector g_1. For $j = 1$, this is obvious. Suppose that this holds for $j = i-1$. Since $x_i = x_{i-1} + \alpha_{i-1} s_{i-1}$ implies $g_i = g_{i-1} + \alpha_{i-1} G s_{i-1}$ (property of the quadratic function (Q)) and since $g_{i-1} \in \mathcal{K}_{i-1}$ and $Gs_{i-1} \in \mathrm{span}(Gg_1, G^2 g_1, \ldots, G^{i-1} g_1) \subset \mathcal{K}_i$ (inductive assumption), we obtain $g_i \in \mathcal{K}_i$. Furthermore, since $s_i = -g_i + \beta_{i-1} s_{i-1}$ (CG) and since $s_{i-1} \in \mathcal{K}_{i-1} \subset \mathcal{K}_i$ (inductive assumption) and $g_i \in \mathcal{K}_i$ (proved inclusion), we obtain $s_i \in \mathcal{K}_i$.

(b) Using (a) we can write

$$\begin{aligned} x_{i+1} - x^* &= x_1 - x^* + \sum_{j=1}^{i} \alpha_j s_j = x_1 - x^* + P_{i-1}^*(G) g_1 = \\ &= x_1 - x^* + P_{i-1}^*(G) G (x_1 - x^*) = (I + G P_{i-1}^*(G))(x_1 - x^*) \end{aligned}$$

where $P^*_{i-1}(G)$ is a polynomial of degree $i-1$ in G. Denote $\overline{P}^*_i = I + GP^*_{i-1}$ (so that $\overline{P}^*_i$ has degree i and $\overline{P}^*_i(0) = 1$). Since the proof of Theorem 10 implies that $s_j^T g_{i+1} = 0 \; \forall 1 \le j \le i$, then

$$x_{i+1} = x^* + \overline{P}^*_i(G)(x_1 - x^*) = \arg \min_{x = x^* + \overline{P}_i(G)(x_1 - x^*)} Q(x)$$

holds, so that

$$\begin{aligned} Q(x_{i+1}) - Q(x^*) &= \frac{1}{2}(x_{i+1} - x^*)^T G(x_{i+1} - x^*) = \frac{1}{2}(x_1 - x^*)^T \overline{P}^*_i(G) G \overline{P}^*_i(G)(x_1 - x^*) \le \\ &\le \frac{1}{2}(x_1 - x^*)^T \overline{P}_i(G) G \overline{P}_i(G)(x_1 - x^*) \end{aligned}$$

for an arbitrary polynomial $\overline{P}_i$ of degree i, such that $\overline{P}_i(0) = 0$. Let λ_k and v_k, $1 \le k \le n$, be the (nonnegative) eigenvalues and (orthonormal) eigenvectors of the matrix G, respectively, and let

$$x_1 - x^* = \sum_{k=1}^n \gamma_k v_k$$

Then

$$Q(x_1) - Q(x^*) = \frac{1}{2}(x_1 - x^*)^T G(x_1 - x^*) = \frac{1}{2} \left(\sum_{k=1}^n \gamma_k v_k \right)^T G \left(\sum_{k=1}^n \gamma_k v_k \right) = \frac{1}{2} \sum_{k=1}^n \gamma_k^2 \lambda_k$$

and

$$\begin{aligned} Q(x_{i+1}) - Q(x^*) &\le \frac{1}{2}(x_1 - x^*)^T \overline{P}_i(G) G \overline{P}_i(G)(x_1 - x^*) = \\ &= \frac{1}{2} \left(\sum_{k=1}^n \overline{P}_i(\lambda_k) \gamma_k v_k \right)^T G \left(\sum_{k=1}^n \overline{P}_i(\lambda_k) \gamma_k v_k \right) = \\ &= \frac{1}{2} \sum_{k=1}^n \overline{P}_i^2(\lambda_k) \gamma_k^2 \lambda_k \le \frac{1}{2} \max_{1 \le k \le n} \overline{P}_i^2(\lambda_k) \sum_{k=1}^n \gamma_k^2 \lambda_k. \end{aligned}$$

After division we obtain

$$\frac{Q(x_{i+1}) - Q(x^*)}{Q(x_1) - Q(x^*)} \le \max_{1 \le k \le n} \overline{P}_i^2(\lambda_k).$$

Theorem 12 Let the assumptions of Theorem 10 be satisfied. Let $g_i \ne 0$ for some index $1 \le i \le n$. Then

$$\frac{Q(x_{i+1}) - Q(x^*)}{Q(x_1) - Q(x^*)} \le 4 \left(\frac{\sqrt{\kappa(G)} - 1}{\sqrt{\kappa(G)} + 1} \right)^{2i}.$$

Proof Lemma 4 implies

$$\frac{Q(x_{i+1}) - Q(x^*)}{Q(x_1) - Q(x^*)} \le \max_{1 \le k \le n} (\overline{P}_i(\lambda_k))^2$$

for an arbitrary polynomial $\overline{P}_i(\lambda)$ of degree at most i such that $\overline{P}_i(0) = 1$. Choose the polynomial $\overline{P}_i(\lambda)$ in such a way to minimize the value

$$\max_{\lambda_1 \le \lambda \le \lambda_n} |\overline{P}_i(\lambda)|.$$

This property has the Chebyshev polynomial transformed on the interval $\lambda_1 \le \lambda \le \lambda_n$ and normalized in such a way to reach the value 1 for $\lambda = 0$, i.e. we set

$$\overline{P}_i(\lambda) = \frac{T_i\left(\frac{\lambda_n+\lambda_1-2\lambda}{\lambda_n-\lambda_1}\right)}{T_i\left(\frac{\lambda_n+\lambda_1}{\lambda_n-\lambda_1}\right)},$$

where $T_i(\xi) = \cos(i \arccos \xi)$ pro $|\xi| \le 1$ and $T_i(\xi) = ((\xi + \sqrt{\xi^2-1})^i + (\xi - \sqrt{\xi^2-1})^i)/2$ for $|\xi| \ge 1$. Since $|T_i(\xi)| \le 1$ for $|\xi| \le 1$, we get

$$\max_{\lambda_1 \le \lambda \le \lambda_n} |\overline{P}_i(\lambda)| \le 1/T_i\left(\frac{\lambda_n+\lambda_1}{\lambda_n-\lambda_1}\right).$$

It remains to calculate the value on the rigt hand side of the last inequality. Denote $\xi = (\lambda_n + \lambda_1)/(\lambda_n - \lambda_1)$. Obviously $|\xi| \ge 1$, so that

$$\begin{aligned} T_i(\xi) &= \frac{1}{2}((\xi + \sqrt{\xi^2-1})^i + (\xi - \sqrt{\xi^2-1})^i) \ge \frac{1}{2}(\xi + \sqrt{\xi^2-1})^i = \\ &= \frac{1}{2}\frac{1}{2^i}(\sqrt{\xi+1} + \sqrt{\xi-1})^{2i}, \end{aligned}$$

since

$$(\sqrt{\xi+1} + \sqrt{\xi-1})^2 = 2(\xi + \sqrt{\xi^2-1})$$

After substituting the value ξ, we obtain

$$\begin{aligned} T_i\left(\frac{\lambda_n+\lambda_1}{\lambda_n-\lambda_1}\right) &\ge \frac{1}{2}\left(\sqrt{\frac{\lambda_n}{\lambda_n-\lambda_1}} + \sqrt{\frac{\lambda_1}{\lambda_n-\lambda_1}}\right)^{2i} = \frac{1}{2}\left(\frac{(\sqrt{\lambda_n}+\sqrt{\lambda_1})^2}{\lambda_n-\lambda_1}\right)^i = \\ &= \frac{1}{2}\left(\frac{\sqrt{\lambda_n}+\sqrt{\lambda_1}}{\sqrt{\lambda_n}-\sqrt{\lambda_1}}\right)^i. \end{aligned}$$

Therefore, it holds

$$\begin{aligned} \frac{Q(x_{i+1}) - Q(x^*)}{Q(x_1) - Q(x^*)} &\le \left(\max_{1 \le k \le n} |\overline{P}_i(\lambda_k)|\right)^2 \le \left(\frac{1}{T_i\left(\frac{\lambda_n+\lambda_1}{\lambda_n-\lambda_1}\right)}\right)^2 \le \\ &\le 4\left(\frac{\sqrt{\lambda_n}-\sqrt{\lambda_1}}{\sqrt{\lambda_n}+\sqrt{\lambda_1}}\right)^{2i} = 4\left(\frac{\sqrt{\kappa(G)}-1}{\sqrt{\kappa(G)}+1}\right)^{2i}. \end{aligned}$$

The following proposition is a corollary of Proposition 3 (with results obtained in its proof) and Theorem 12.

Proposition 4 (Asymptotic rate of convergence) Let the assumptions of Theorem 11 be satisfied and let the line search be asymptotically perfect. Let the function $F : R^n \to R$ satisfy in addition the condition (F5). Then

$$\limsup_{i \to \infty} \| x_i - x^* \|^{\frac{1}{i}} \le \frac{\sqrt{\kappa(G^*)}-1}{\sqrt{\kappa(G^*)}+1}.$$

Remark 22 The asymptotic rate $(\sqrt{\kappa}-1)/(\sqrt{\kappa}+1)$ is much better than the asymptotic rate $(\kappa - 1)/(\kappa+1)$ of the steepest descent method, as demonstrated in the following table, which contains the number of iterations required for obtaining the solution with precision ε.

	SD	CG
$\kappa = 10^2$, $\varepsilon = 10^{-4}$	460	45
$\kappa = 10^4$, $\varepsilon = 10^{-6}$	69077	690
$\kappa = 10^6$, $\varepsilon = 10^{-8}$	9210340	9210

Now we state several comments concerning the implementation of conjugate gradient methods.

1) Conjugate gradient methods require some modifications of the line search algorithm:

a) The initial stepsize

$$\alpha = \min\left(1, \frac{2(F_i - F_{i-1})}{s_i^T g_i}, \frac{20(\underline{F} - F_i)}{s_i^T g_i}\right)$$

is used, where $\underline{F}$ is a lower bound for the minimum value of the function F.

b) The strong Wolfe condition (S3a)

$$|s_i^T g_{i+1}| \le \varepsilon_2 |s_i^T g_i|,$$

where $\varepsilon_2 = 10^{-1}$, has to be used instead of the weak Wolfe condition (S3b). Therefore, Algoritm 1 has to be modified in such a way that the condition (S3b) remains unchanged but we add, to (S2), the extra condition

$$s_i^T g_{i+1} \le \varepsilon_2 |s_i^T g_i|,$$

which is a part of the condition (S3a).

2) It is advantageous to use scaling of conjugate gradient methods (the simplest preconditioning). Instead of

$$-s_{i+1} = g_{i+1} - \beta_i s_i,$$

we use the formula

$$-s_{i+1} = \gamma_{i+1}(g_{i+1} - \beta_i s_i),$$

where

$$\gamma_{i+1} = \frac{y_i^T d_i}{y_i^T y_i}$$

($y_i = g_{i+1} - g_i, d_i = x_{i+1} - x_i = \alpha_i s_i$) In this case, the formulas defining the value of the parameter β_i have to be modified in such a way that

$$\beta_i = \frac{y_i^T g_{i+1}}{y_i^T s_i}, \tag{CGa}$$

$$\beta_i = \frac{1}{\gamma_i} \frac{y_i^T g_{i+1}}{g_i^T g_i}, \tag{CGb}$$

$$\beta_i = \frac{1}{\gamma_i} \frac{g_{i+1}^T g_{i+1}}{g_i^T g_i}, \tag{CGc}$$

Values of the parameter γ_{i+1} have to be kept in the interval $(0.005 \le \gamma_{i+1} \le 200)$, say.

3) It is advantageous to use controlled restarts:

a) Classic way. The iterative process is restarted after n iterations or if $\beta_i \le 0$.

b) Comparison with (CGc). The iterative process is restarted if

$$\eta_1 \beta_i^{FR} \le \beta_i \le \eta_2 \beta_i^{FR}$$

does not hold, where $\beta_i^{FR} = g_{i+1}^T g_{i+1} / g_i^T g_i$ and $\eta_1 \approx 0.1 - 0.3$ and $\eta_2 \approx 1.1 - 1.3$.

c) Conjugacy test. The iterative process is restarted if

$$s_{i+1}^T y_i \geq \eta_3 \| s_{i+1} \| \| y_i \|,$$

where $\eta_3 \approx 0.04 - 0.05$.

d) Orthogonality test. The iterative process is restarted if

$$g_{i+1}^T g_i \geq \eta_4 \| g_{i+1} \| \| g_i \|,$$

where $\eta_4 \approx 0.4 - 0.5$.

An implementation of conjugate gradient methods can be roughly described by the following algorithm (numbers and letters are related to above comments):

Algorithm 2 (CG) Data $\varepsilon_1 = 10^{-4}$, $\varepsilon_2 = 10^{-1}$, $\eta_1 = 0.3$, $\eta_2 = 1.2$, $\eta_3 = 0.05$, $\eta_4 = 0.46$, $\underline{\varepsilon} > 0$.

Step 1 Choose an initial point $x_1 \in R^n$, compute $F_1 = F(x_1)$, $g_1 = g(x_1)$ and set $i = 1$.

Step 2 If $\| g_i \| \leq \underline{\varepsilon}$, then terminate the computations. Otherwise determine the coefficient β_{i-1} by one of the formulas (CGa), (CGb), (CGc) and make a decision about restart using one of the strategies 3a, 3b, 3c, 3d. Choose scaling coefficient γ_i and determine the direction vector s_i by 2.

Step 3 Select the stepsize α_i using Algorithm 1 modified by 1a a 1b. Set $x_{i+1} = x_i + \alpha_i s_i$, compute $F_{i+1} = F(x_{i+1})$, $g_{i+1} = g(x_{i+1})$, increase i about 1 and go to step 2.

The following table gives a comparison of individual conjugate gradient methods by means of 24 test problems with 100 variables (the total numbers of iterations NIT and function evaluations NFV are shown together with the total computational time in seconds).

method	NIT-NFV	fail	time
CGc + 3a	10152 - 18772	1	2:13.90
CGc + 3c	4508 - 8952	-	1:01.84
CGc + 3d	4140 - 8397	-	1:00.91
CGb + 3a	7568 - 14917	1	1:48.20
CGb + 3b	5690 - 11248	-	1:09.10
CGa + 3a	10357 - 20423	1	1:31.51
CGa + 3b	5140 - 10111	-	1:07.18

2.3 Variable metric methods

Definition 21 We say that the basic optimization method is a variable metric method if

$$s_i = -H_i g_i \qquad \forall i \in N, \tag{VM1}$$

where H_i, $i \in N$, are symmetric positive definite (SPD) matrices generated by the recurrence relation

$$H_{i+1} = \gamma_i (H_i + U_i M_i U_i^T), \tag{VM2}$$

where $U_i \in R^{n \times 2}$, $M_i \in R^{2 \times 2}$ (symmetric) and $\gamma_i > 0$, and satisfying the condition

$$H_{i+1} y_i = \rho_i d_i, \tag{VM3}$$

where $y_i = g_{i+1} - g_i$, $d_i = x_{i+1} - x_i = \alpha_i s_i$ and $\rho_i > 0$.

Remark 23 The matrix H_{i+1} is obtained from the matrix H_i by an update with rank at most 2. The most efficient variable metric methods belong to the Broyden class, which is characterized by the choice $U_i = [d_i, H_i y_i]$.

Theorem 13 (Quadratic termination) Let x_i, $i \in N$, be a sequence generated by a variable metric method from the Broyden class with perfect line search ($s_i^T g_{i+1} = 0$ holds $\forall i \in N$) applied to the strict convex quadratic function (Q). Then an index $k \le n$ exists, such that $g_{k+1} = 0$ and $x_{k+1} = x^*$.

Proof Suppose that $g_i \ne 0$ $\forall 1 \le i \le n$. We prove by induction that $s_i \ne 0$ and $\alpha_i \ne 0$ $\forall 1 \le i \le n$ and that it holds

$$(\alpha) \qquad H_i y_j = \lambda_i^j d_j \qquad \forall 1 \le j < i \le n+1,$$

$$(\beta) \qquad s_j^T g_i = 0 \qquad \forall 1 \le j < i \le n+1,$$

$$(\gamma) \qquad s_j^T G s_i = 0 \qquad \forall 1 \le j < i \le n.$$

From (VM1) and (γ) it follows that s_i, $1 \le i \le n$, are nonzero and mutually conjugate (G-orthogonal), hence linearly independent, so that (β) necessarily implies $g_{n+1} = 0$. For $i = 1$, $s_1^T g_1 = -s_1^T H_1 s_1 < 0$ (H_1 is SPD) holds, so that $s_1 \ne 0$ and $\alpha_1 \ne 0$ and nothing else has to be proved. The induction steps:

(a) Let $i \le n$. The inductive assumption (γ) and Q imply $d_i^T y_j = d_i^T G d_j = \alpha_i \alpha_j s_i^T G s_j = 0$ and (α) in addition implies $y_i^T H_i y_j = \lambda_i^j y_i^T d_j = \lambda_i^j d_i^T G d_j = 0$, so that $U_i^T y_j = 0$ $\forall 1 \le j < i$. Therefore, by (VM2) and (α) it holds

$$H_{i+1} y_j = \gamma_i (H_i y_j + U_i^T M_i U_i^T y_j) = \gamma_i H_i y_j = \gamma_i \lambda_i^j d_j \triangleq \lambda_{i+1}^j d_j$$

$\forall 1 \le j < i$. Using (VM3) we get $H_{i+1} y_i = \rho_i d_i \triangleq \lambda_{i+1}^i d_i$, so that $H_{i+1} y_j = \lambda_{i+1}^j d_i$ $\forall 1 \le j \le i$.

(b) Let $i \le n$. Using (γ) and (Q), we obtain $s_j^T g_{i+1} = s_j^T g_i + s_j^T y_i = s_j^T g_i + \alpha_i s_j^T G s_i = 0$ $\forall 1 \le j < i$. The perfect line search implies $s_i^T g_{i+1} = 0$, so that finally $s_j^T g_{i+1} = 0$ $\forall 1 \le j \le i$.

(c) Formula (VM1) implies $g_{i+1}^T s_{i+1} = -g_{i+1}^T H_{i+1} g_{i+1} < 0$ so that $s_{i+1} \ne 0$ and $\alpha_{i+1} \ne 0$. Using (VM1), (Q), (a), (b) we get

$$s_j^T G s_{i+1} = -\frac{1}{\alpha_j} y_j^T H_{i+1} g_{i+1} = -\frac{\lambda_{i+1}^j}{\alpha_j} d_j^T g_{i+1} = -\lambda_{i+1}^j s_j^T g_{i+1} = 0$$

$\forall 1 \le j \le i$.

Theorem 14 (Approximation of the Hessian matrix). Let the assumptions of Theorem 13 be satisfied with $\gamma_i = 1$ and $\rho_i = 1$ $\forall i \in N$. Then $H_{n+1} = G^{-1}$.

Proof From the proof of Theorem 13 it follows that

$$H_{n+1} y_j = d_j \qquad \forall 1 \le j \le n$$

and that the vectors d_j a $y_j = G d_j$, $1 \le j \le n$, are linearly independent. For this reason, $H_{n+1} = G^{-1}$ must hold.

Remark 24 Now we will investigate the update (VM3). For simplifying the notation we omit the index i and replace the index $i+1$ by $+$.

Theorem 15 Let $H_+ = \gamma(H + UMU^T)$, where $U = [d, Hy]$ is a matrix of rank 2. Then $H_+ y = \rho d$ holds if and only if

$$M = \begin{bmatrix} \frac{1}{b}\left(\eta \frac{a}{b} + \frac{\rho}{\gamma}\right), & -\frac{\eta}{b} \\ -\frac{\eta}{b}, & \frac{\eta - 1}{a} \end{bmatrix},$$

where η is a free parameter and where

$$a = y^T H y, \qquad b = y^T d, \qquad c = d^T H^{-1} d.$$

Proof By (VM2) and (VM3) it follows that

$$\begin{aligned} H_+ y &= \gamma \left(Hy + [d, Hy] \begin{bmatrix} m_1 & m_2 \\ m_2 & m_3 \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} \right) = \\ &= \gamma (Hy + (m_1 b + m_2 a) d + (m_2 b + m_3 a) Hy) = \rho d \end{aligned}$$

so that neccessarily

$$m_1 b + m_2 a = \rho / \gamma,$$

$$m_2 b + m_3 a = -1.$$

One parameter is overdetermined. Therefore, we choose $m_2 = -\eta / b$ and the remaining two parameters m_1, m_3 are determined as a solution of the above system. In this way we obtain the matrix M introduced in Theorem 15.

Remark 25 Using the expression of M and U, the relation $H_+ = \gamma (H + UMU^T)$ gives

$$H_+ = \gamma \left(H + \frac{\rho}{\gamma} \frac{1}{b} dd^T - \frac{1}{a} Hy (Hy)^T + \frac{\eta}{a} \left(\frac{a}{b} d - Hy \right) \left(\frac{a}{b} d - Hy \right)^T \right), \tag{H}$$

the (scaled) Broyden class. The most popular members of this class are obtained if we substitute $\eta = 0$, the (scaled) DFP method:

$$H_+ = \gamma \left(H + \frac{\rho}{\gamma} \frac{1}{b} dd^T - \frac{1}{a} Hy (Hy)^T \right) \tag{HD}$$

or $\eta = 1$, the (scaled) BFGS method:

$$H_+ = \gamma \left(H + \left(\frac{a}{b} + \frac{\rho}{\gamma} \right) \frac{1}{b} dd^T - \frac{1}{b} (Hyd^T + d (Hy)^T) \right), \tag{HB}$$

or $\eta = (\rho/\gamma)/(\rho/\gamma - a/b)$, the (scaled) symmetric rank 1 method):

$$H_+ = \gamma \left(H + \frac{1}{(\rho/\gamma) b - a} \left(\frac{\rho}{\gamma} d - Hy \right) \left(\frac{\rho}{\gamma} d - Hy \right)^T \right) \tag{HR}$$

From these three particular methods, only the BFGS method can be used in its basic form. The DFP method requires either accurate line search or scaling in every iteration. Otherwise, it converges very slowly (more detailed reasons are given in Remark 34). The symmetric rank 1 method does not satisfy, in general, the condition for positive definiteness of the matrix H_+ (this condition is given in Theorem 16), so that global convergence can be lost by violation of the condition (S1a). The symmetric rank 1 method is frequently combined with the BFGS method.

Lemma 5 Let H be a SPD matrix, $a > 0$, $b > 0$, $c > 0$, $ac - b^2 > 0$ and let (H) hold, where $\gamma > 0$ and $\rho > 0$. Then the matrix $(1/\gamma) H^{-\frac{1}{2}} H_+ H^{-\frac{1}{2}}$ has $n-2$ unit eigenvalues and the remaining two eigenvalues area solutions of the quadratic equation

$$\lambda^2 - p\lambda + q = 0,$$

where

$$p = \frac{1}{b^2} (\eta (ac - b^2) + b^2) + \frac{\rho}{\gamma} \frac{c}{b},$$

$$q = \frac{\rho}{\gamma} \frac{1}{ab} (\eta (ac - b^2) + b^2).$$

Proof Using (VM2), we obtain

$$\frac{1}{\gamma}H^{-\frac{1}{2}}H_+H^{-\frac{1}{2}} = I + H^{-\frac{1}{2}}UMU^TH^{-\frac{1}{2}}.$$

This matrix has $n-2$ unit eigenvalues corresponding to $n-2$ eigenvectors orthogonal to $H^{-\frac{1}{2}}U$. The remaining two eigenvectors can be expressed in the form $H^{-\frac{1}{2}}Uz$, so that the corresponding eigenvalues are solutions of the equation

$$H^{-\frac{1}{2}}U(I + MU^TH^{-1}U)z = \lambda H^{-\frac{1}{2}}Uz$$

or (after multiplying by $(U^TH^{-1}U)^{-1}U^TH^{-\frac{1}{2}}$)

$$((1-\lambda)I + MU^TH^{-1}U)z = 0.$$

Substituting M from Theorem 15 and letting

$$U^TH^{-1}U = \begin{bmatrix} c & b \\ b & a \end{bmatrix}$$

we can write

$$\begin{aligned}\det((1-\lambda)I + MU^TH^{-1}U) &= \det\left(\begin{bmatrix} 1-\lambda, & 0 \\ 0, & 1-\lambda \end{bmatrix} + \right. \\ &\left. + \begin{bmatrix} \frac{1}{b}\left(\eta\frac{a}{b} + \frac{\rho}{\gamma}\right) & -\frac{\eta}{b} \\ -\frac{\eta}{b}, & -\frac{\eta-1}{a} \end{bmatrix}\begin{bmatrix} c, & b \\ b, & a \end{bmatrix}\right) = 0,\end{aligned}$$

which after straightforward manipulations gives $\lambda^2 - p\lambda + q$ with coefficients given in Lemma 5.

Theorem 16 Let the assumptions of Lemma 5 be satisfied. Then H_+ is SPD if and only if $\eta(ac - b^2) + b^2 > 0$.

Proof It is necessary to find a condition guaranteeing for the equation $\lambda^2 - p\lambda + q$, with coefficients given in Lemma 5, to have positive roots. Denote these roots by λ_1 and λ_2. Then $\lambda_1 + \lambda_2 = p$ and $\lambda_1\lambda_2 = q$ so that $\lambda_1 > 0$ and $\lambda_2 > 0$ if and only if $p > 0$ and $q > 0$. The definition of p and q implies

$$p = \frac{a}{b}\frac{\gamma}{\rho}q + \frac{\rho}{\gamma}\frac{c}{b}.$$

If we assume $a > 0$, $b > 0$, $c > 0$, $\gamma > 0$, $\rho > 0$, then $p > 0$ whenever $q > 0$. From $q > 0$ we get the condition $\eta(ac - b^2) + b^2 > 0$.

Remark 26 The last inequality can be written in the form $\eta > \eta^*$, where

$$\eta^* = -\frac{b^2}{ac - b^2} < 0.$$

The conditions $a > 0$, $c > 0$ are satisfied, if the matrix H is SPD and $d \neq 0$, $y \neq 0$. The condition $b > 0$ is satisfied if the stepsize selection satisfies the condition (S3b), since

$$y^Td = \alpha(g_+ - g)^Ts \geq \alpha(\varepsilon_2 - 1)g^Ts > 0$$

in this case. If $b \leq 0$, then the matrix H^+ is not SPD for any values of parameters γ, ρ and η.

Theorem 17 (Update of the matrix $B = H^{-1}$). Let the assumptions of Lemma 5 be satisfied. let $B = H^{-1}$ and let H_+ be a matrix determined by the update (H). Let $B_+ = H_+^{-1}$. Then

$$B_+ = \frac{1}{\gamma}\left(B + \frac{\gamma}{\rho}\frac{1}{b}yy^T - \frac{1}{c}Bd(Bd)^T + \frac{\beta}{c}\left(\frac{c}{b}y - Bd\right)\left(\frac{c}{b}y - Bd\right)^T\right) \qquad \text{(B)},$$

where

$$\beta\eta(ac - b^2) + (\beta + \eta)b^2 = b^2.$$

Proof By inversion of the relation $H_+ = \gamma(H + UMU^T)$, we obtain

$$B_+ = \frac{1}{\gamma}(B - BU(M^{-1} + U^T BU)^{-1} U^T B) \triangleq \frac{1}{\gamma}(B + BUKU^T B)$$

(Woodbury formula), where $K \in R^{2\times 2}$. Since (VM3) implies $H_+ y = \rho d$, then $B_+ d = (1/\rho)y$ has to hold or

$$B_+ d = \frac{1}{\gamma}\left(Bd + [Bd, y]\begin{bmatrix} k_1 & k_2 \\ k_2 & k_3 \end{bmatrix}\begin{bmatrix} c \\ b \end{bmatrix}\right) = \frac{1}{\gamma}(Bd + (k_1 c + k_2 b)Bd + (k_2 c + k_3 b)y) = \frac{1}{\rho}y,$$

so that necessarily

$$k_1 c + k_2 b = -1$$

$$k_2 c + k_3 b = \gamma/\rho.$$

Choose $k_2 = -\beta/b$ and determine the remaining two elements k_1, k_3 as a solution of the above system. In this way we obtain

$$K = \begin{bmatrix} \frac{\beta-1}{c}, & -\frac{\beta}{b} \\ -\frac{\beta}{b} & \frac{1}{b}\left(\beta\frac{c}{b} + \frac{\gamma}{\rho}\right) \end{bmatrix},$$

which after a substitution into $B_+ = (1/\gamma)(B + BUKU^T B)$ gives (B). The relation between β and η can be obtained from the equality

$$K = -(M^{-1} + U^T BU)^{-1}.$$

(proof omitted).

Remark 27 (Duality) The relation (B) can be obtained fron the relation (H) by the replacements $\gamma \to 1/\gamma$, $\rho \to 1/\rho$, $a \to c$, $c \to a$, $d \to y$, $y \to d$, $H \to B$, $\eta \to \beta$. The methods DFP and BFGS are dual to each other. The method DFP corresponds to $\beta = 1$:

$$B_+ = \frac{1}{\gamma}\left(B + (\frac{c}{b} + \frac{\gamma}{\rho})\frac{1}{b}yy^T - \frac{1}{b}(Bdy^T + y(Bd)^T)\right). \tag{BD}$$

The method BFGS corresponds to $\beta = 0$:

$$B_+ = \frac{1}{\gamma}\left(B + \frac{\gamma}{\rho}\frac{1}{b}yy^T - \frac{1}{c}Bd(bd)^T\right). \tag{BB}$$

The symmetric rank 1 method is self dual and it corresponds to $\beta = (\gamma/\rho)/(\gamma/\rho - c/b)$:

$$B_+ = \frac{1}{\gamma}\left(B + \frac{1}{(\gamma/\rho)b - c}\left(\frac{\gamma}{\rho}y - Bd\right)\left(\frac{\gamma}{\rho}y - Bd\right)^T\right). \tag{BR}$$

Remark 28 The matrix B_+ is SPD if and only if $\beta > \beta^*$, where

$$\beta^* = -\frac{b^2}{ac - b^2} < 0.$$

Theorem 18 Let the assumptions of Lemma 5 be satisfied and let

$$\begin{aligned} q &= \frac{\rho}{\gamma}\frac{1}{ab}\left(\eta(ac - b^2) + b^2\right) \ \geq 0, \\ m &= \frac{1}{ab}\left(\eta\left(\frac{a}{b} - \frac{\rho}{\gamma}\right) + \frac{\rho}{\gamma}\right) \ \geq 0. \end{aligned}$$

Let $H = SS^T$ and let H_+ be a matrix determined by the update (H). Denote $\tilde{U} = S^{-1}U = [S^{-1}d, S^T y] \triangleq [\tilde{d}, \tilde{y}]$. Then $H_+ = S_+ S_+^T$, where

$$S_+ = \sqrt{\gamma} S(I + \tilde{U} uv^T \tilde{U}^T) \tag{S}$$

and

$$uv^T = \frac{1}{\lambda} \begin{bmatrix} \frac{\rho}{\gamma} + a\sqrt{m} \\ -1 - b\sqrt{m} \end{bmatrix} \begin{bmatrix} \sqrt{q} - b\sqrt{m} \\ -\frac{\gamma}{\rho}\sqrt{q} + c\sqrt{m} \end{bmatrix}^T,$$

where

$$\lambda = \left(\frac{\rho}{\gamma}c - b\right) + \left(b - \frac{\gamma}{\rho}a\right)\sqrt{q} + (ac - b^2)\sqrt{m}$$

Proof (a) Substituting the relation (S) into the expression $H_+ = S_+ S_+^T$, we obtain

$$H_+ = \gamma(I + Uuv^T U^T H^{-1}) H (I + Uuv^T U^T H^{-1})^T.$$

Comparing this relation with the formula $H_+ = \gamma(H + UMU^T)$, we can write

$$M = uv^T + vu^T + uv^T U^T H^{-1} U vu^T = [u, v] \begin{bmatrix} v^T U^T H^{-1} U v & , & 1 \\ 1 & , & 0 \end{bmatrix} \begin{bmatrix} u^T \\ v^T \end{bmatrix}$$

and using the rule for the product of determinants we obtain

$$(u_1 v_2 - v_1 u_2)^2 = (\det[u, v])^2 = -\det M = m,$$

since from the expression of the matrix M (Theorem 15) it follows that

$$\det M = -\frac{1}{ab}\left(\eta\left(\frac{a}{b} - \frac{\rho}{\gamma}\right) + \frac{\rho}{\gamma}\right) = -m.$$

Using this result, we can write

$$uv^T - vu^T = \begin{bmatrix} 0 & , & u_1 v_2 - v_1 u_2 \\ v_1 u_2 - u_1 v_2 & , & 0 \end{bmatrix} = \begin{bmatrix} 0, & +\sqrt{m} \\ -\sqrt{m}, & 0 \end{bmatrix}.$$

(b) Substituting the relation (S) into the expression $H_+ = S_+ S_+^T$, we get

$$\frac{1}{\gamma} H_+ = S(I + \tilde{U} uv^T \tilde{U}^T)(I + \tilde{U} vu^T \tilde{U}^T) S^T.$$

Since we know, from proof of lemma 5, that $\det(H_+/\gamma) = q \det H$, we can write

$$\det(I + \tilde{U} uv^T \tilde{U}^T) = \sqrt{q},$$

so that using the Sherman-Morrison formula we obtain

$$(I + \tilde{U} uv^T \tilde{U}^T)^{-1} = I - \frac{1}{\sqrt{q}} \tilde{U} uv^T \tilde{U}^T$$

and the condition $H_+ y = \rho d$ can be written in the form

$$(I + \tilde{U} vu^T \tilde{U}^T)\tilde{y} = \frac{\rho}{\gamma}\left(I - \frac{1}{\sqrt{q}} \tilde{U} uv^T \tilde{U}^T\right)\tilde{d}.$$

Multiplying this equation from left by the matrix $\tilde{U}^T$ and taking into account that

$$\tilde{U}^T \tilde{U} = \begin{bmatrix} \tilde{d}^T \tilde{d} & \tilde{d}^T \tilde{y} \\ \tilde{y}^T \tilde{d} & \tilde{y}^T \tilde{y} \end{bmatrix} = \begin{bmatrix} c, & b \\ b, & a \end{bmatrix},$$

we obtain

$$\begin{bmatrix} b \\ a \end{bmatrix} + \begin{bmatrix} c, & b \\ b, & a \end{bmatrix} vu^T \begin{bmatrix} b \\ a \end{bmatrix} = \frac{\rho}{\gamma} \left(\begin{bmatrix} c \\ b \end{bmatrix} - \frac{1}{\sqrt{q}} \begin{bmatrix} c, & b \\ b, & a \end{bmatrix} uv^T \begin{bmatrix} c \\ b \end{bmatrix} \right),$$

which after straighforward manipulations gives

$$vu^T \begin{bmatrix} b \\ a \end{bmatrix} + uv^T \frac{1}{\sqrt{q}} \frac{\rho}{\gamma} \begin{bmatrix} c \\ b \end{bmatrix} = \frac{1}{ac - b^2} \begin{bmatrix} a, & -b \\ -b, & c \end{bmatrix} \left(\frac{\rho}{\gamma} \begin{bmatrix} c \\ b \end{bmatrix} - \begin{bmatrix} b \\ a \end{bmatrix} \right) = \begin{bmatrix} \frac{\rho}{\gamma} \\ -1 \end{bmatrix}.$$

If we now use (a), we get

$$uv^T \left(\begin{bmatrix} b \\ a \end{bmatrix} + \frac{1}{\sqrt{q}} \frac{\rho}{\gamma} \begin{bmatrix} c \\ b \end{bmatrix} \right) = \begin{bmatrix} \frac{\rho}{\gamma} + a\sqrt{m} \\ -1 - b\sqrt{m} \end{bmatrix}.$$

From this expression, it can be seen thet the vector $u \in R^2$ is a multiple of the vector appearing on the right hand side of the last equality. Since the norm of this vector can be chosen arbitrarily, we can set

$$u = \begin{bmatrix} \frac{\rho}{\gamma} + a\sqrt{m} \\ -1 - b\sqrt{m} \end{bmatrix}.$$

Then the vector $v \in R^2$ has to satisfy the equation

$$v_1 \left(b + \frac{1}{\sqrt{q}} \frac{\rho}{\gamma} c \right) + v_2 \left(a + \frac{1}{\sqrt{q}} \frac{\rho}{\gamma} b \right) = 1$$

and (a) implies

$$v_1 \left(1 + b\sqrt{m} \right) + v_2 \left(\frac{\rho}{\gamma} + a\sqrt{m} \right) = \sqrt{m}.$$

The following vector is a solution to these two equations

$$v = \frac{1}{\lambda} \begin{bmatrix} \sqrt{q} - b\sqrt{m} \\ -\frac{\gamma}{\rho}\sqrt{q} + c\sqrt{m} \end{bmatrix}.$$

Here

$$\lambda = \left(\frac{\rho}{\gamma} c - b \right) + \left(b - \frac{\gamma}{\rho} a \right) \sqrt{q} + (ac - b^2)\sqrt{m}.$$

Remark 29 The formula (S) can be used, instead of the relation $s = -Hg$, in such a way that

$$\tilde{g} = S^T g, \qquad \tilde{s} = -\tilde{g}, \qquad s = S\tilde{s}.$$

Since $d = \alpha s$, we can set $\tilde{d} = \alpha \tilde{s}$ and $\tilde{y} = S^T(g_+ - g)$. The matrix S is then updated by the formula

$$S_+ = \sqrt{\gamma} S (I + \tilde{u}\tilde{v}^T),$$

where $\tilde{u} = u_1 \tilde{d} + u_2 \tilde{y}$ and $\tilde{v} = v_1 \tilde{d} + v_2 \tilde{y}$ (the numbers u_1, u_2 and v_1, v_2 are determined by Theorem 18). This way is used especially in the case when the matrix S is rectangular, e.g. in linearly constrained optimization when the matrix S has more rows than columns.

Remark 30 The DFP methods uses the value $\eta = 0$, so that $q = \rho b/(\gamma a)$ and $m = \rho/(\gamma a b)$. After substituting these values into (S) we obtain

$$S_+ = \sqrt{\gamma} \left(S + \frac{1}{a} S \left(\sqrt{\frac{\rho a}{\gamma b}} \tilde{d} - \tilde{y} \right) \tilde{y}^T \right) \tag{SD}$$

The BFGS method uses the value $\eta = 1$, so that $q = \rho c/(\gamma b)$ and $m = 1/b^2$. After substituting these values into (S) we obtain

$$S_+ = \sqrt{\gamma}\left(S + \frac{1}{b} S\tilde{d}\left(\sqrt{\frac{\rho b}{\gamma c}}\tilde{d} - \tilde{y}\right)^T\right) \qquad \text{(SB)}.$$

The symmetric rank 1 method uses the value $\eta = (\rho/\gamma)(\rho/\gamma - a/b)$, so that $q = ((\rho/\gamma)c - b)/(b - (\gamma/\rho)a)$ and $m = 0$. After substituting these values into (S) we obtain

$$S_+ = \sqrt{\gamma}\left(S + \frac{\sqrt{q}-1}{\left(\frac{\rho}{\gamma}\right)^2 c - 2\frac{\rho}{\gamma}b + a} S\left(\frac{\rho}{\gamma}\tilde{d} - \tilde{y}\right)\left(\frac{\rho}{\gamma}\tilde{d} - \tilde{y}\right)\right)^T \qquad \text{(SR)}.$$

Remark 31 Theorem 18 implies that $H = SS^T$, where the matrix S is updated by the formula (S). Suppose now that $B = A^T A = (S^{-1})^T S^{-1} = H^{-1}$. The update for the matrix $A = S^{-1}$ can be derived using the Sherman-Morrison formula. Then

$$A_+ = \left(\sqrt{\gamma}S\left(I + \tilde{U}uv^T\tilde{U}^T\right)\right)^{-1} = \frac{1}{\sqrt{\gamma}}\left(I - \frac{1}{\sqrt{q}}\tilde{U}uv^T\tilde{U}^T\right)A \qquad \text{(A)}$$

(see proof of Theorem 18). Fot the DFP method we obtain

$$A_+ = \frac{1}{\sqrt{\gamma}}\left(A + \frac{1}{b}\left(\sqrt{\frac{\gamma b}{\rho a}}\tilde{y} - \tilde{d}\right)\tilde{y}^T A\right) \qquad \text{(AD)}.$$

For the BFGS method we obtain

$$A_+ = \frac{1}{\sqrt{\gamma}}\left(A + \frac{1}{c}\tilde{d}\left(\sqrt{\frac{\gamma c}{\rho b}}\tilde{y} - \tilde{d}\right)^T A\right) \qquad \text{(AB)}$$

For the symmetric rank 1 we obtain

$$A_+ = \frac{1}{\sqrt{\gamma}}\left(A + \frac{1/\sqrt{q}-1}{\left(\frac{\gamma}{\rho}\right)^2 a - 2\frac{\gamma}{\rho}b + c}\left(\frac{\gamma}{\rho}\tilde{y} - \tilde{d}\right)\left(\frac{\gamma}{\rho}\tilde{y} - \tilde{d}\right)^T A\right) \qquad \text{(AR)},$$

where $1/q = ((\gamma/\rho)a - b)/(b - (\rho/\gamma)c)$. Formula (A) is utilized especially for nonlinear least squares problems.

Now we demonstrate an additional property of the variable metric methods.

Theorem 19 Let W be a symmetric positive definite matrix. Then the Frobenius norm $\| W^{-1/2}(\gamma B_+ - B)W^{-1/2} \|_F$ is minimal, on the set of matrices satisfying the condition

$$(\gamma B_+ - B)d = \frac{\gamma}{\rho}y - Bd \triangleq w$$

if and only if

$$B_+ = \frac{1}{\gamma}\left(B + \frac{Wdw^T + w(Wd)^T}{d^TWd} - \frac{w^Td}{d^TWd}\frac{Wd(Wd)^T}{d^TWd}\right).$$

Proof Since the matrix $\gamma B_+ - B$ is symmetric, we can set $\gamma B_+ - B = X + X^T$, where X is a square matrix to be determined. Neccessity will be proved using the Lagrangian function

$$\begin{aligned} L &= \frac{1}{4} \| W^{-1/2}(X + X^T)W^{-1/2} \|_F^2 + u^T(w - (X + X^T)d) = \\ &= \frac{1}{4}\sum_{i=1}^n\sum_{j=1}^n (e_i^T W^{-1/2}(X + X^T)W^{-1/2}e_j)^2 + \sum_{i=1}^n u_i\left(w_i - \sum_{j=1}^n e_i^T(X + X^T)e_j d_j\right), \end{aligned}$$

where $u \in R^n$ is a vector of Lagrange multipliers (e_i, e_j are columns of the unit matrix of order n). Differentiating the Langrangian function we obtain

$$
\begin{aligned}
\frac{\partial L}{\partial X_{kl}} &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(e_i^T W^{-1/2}(X+X^T)W^{-1/2}e_j)(e_i^T W^{-1/2}(e_k e_l^T + e_l e_k^T)W^{-1/2}e_j) - \\
&\quad -\sum_{i=1}^{n} u_i \sum_{j=1}^{n} e_i^T(e_k e_l^T + e_l e_k^T)e_j d_j = \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} e_k^T W^{-1/2}e_i e_i^T W^{-1/2}(X+X^T)W^{-1/2}e_j e_j^T W^{-1/2}e_l + \\
&\quad +\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} e_l^T W^{-1/2}e_i e_i^T W^{-1/2}(X+X^T)W^{-1/2}e_j e_j^T W^{-1/2}e_k + \\
&\quad +\sum_{i=1}^{n}\sum_{j=1}^{n} u_i(e_i^T e_k e_l^T e_j + e_i^T e_l e_k^T e_j)d_j = \\
&= e_k^T W^{-1}(X+X^T)W^{-1}e_l - (u_k d_l + u_l d_k).
\end{aligned}
$$

Therefore, the condition for stationarity of the Langrangian function has the form

$$\gamma B_+ - B = W(ud^T + du^T)W.$$

Substituting this vector into the relation $(\gamma B_+ - B)d = w$, we get after some manipulations

$$(d^T Wd \cdot W + Wd(Wd)^T)u = w.$$

Using the Sherman-Morrison formula, we can express the inverse in the form

$$(d^T Wd \cdot W + Wd(Wd)^T)^{-1} = \frac{1}{d^T Wd}\left(W^{-1} - \frac{dd^T}{2d^T Wd}\right)$$

($d^T Wd \neq 0$ for $d \neq 0$ from positive definiteness of the matrix W). Therefore, it holds

$$u = \frac{1}{d^T Wd}\left(W^{-1} - \frac{dd^T}{2d^T Wd}\right) w,$$

which after substitution into the expression for $\gamma B_+ - B$ gives the statement of the theorem. Sufficiency follows from convexity of a norm.

Remark 32 If we choose the matrix W in such a way that $Wd = y - \lambda Bd$, where λ is selected so that

$$\frac{b(b - \lambda^2(\gamma/\rho)c)}{(b - \lambda c)^2} = \beta$$

then we obtain the update (B). The choice $\lambda = 0$ (or $Wd = y$) corresponds to the DFP method. The choice $\lambda = \sqrt{(\rho/\gamma)(b/c)}$ corresponds to the BFGS method. The choice $\lambda = \rho/\gamma$ corresponds to the symetric rank 1 method. Choosing $W = I$, we obtain the Powell symmetric Broyden method (PSB), which does not belong to the Broyden class

$$B_+ = \frac{1}{\gamma}\left(B + \frac{dw^T + wd^T}{d^T d} - \frac{w^T d}{d^T d}\frac{dd^T}{d^T d}\right).$$

The PSB method does not guarantee positive definiteness of the matrix B_+, so that it may not be globally convergent. Nevertheless, this methods is rather popular, due to its relation to the sparse variable metric updates (Theorem 37).

Lemma 6 Let x_i, $i \in N$ be a sequence generated by a variable metric method from the Broyden class, such that $1 \le \gamma_i \le \overline{\gamma}$, $\underline{\rho} \le \rho_i \le \overline{\rho}$ and $(1-\lambda)\beta_i^* \le \beta_i \le 1-\lambda$, where $0 < \lambda < 1$. Let the function $F : R^n \to R$ satisfy the conditions (F3) and (F4). Then

$$\begin{aligned} Tr(B_{i+1}) &= \frac{1}{\gamma_i}\left(Tr(B_i) + \frac{\gamma_i}{\rho_i}\frac{y_i^T y_i}{y_i^T d_i} + \beta_i \frac{y_i^T y_i}{y_i^T d_i}\frac{d_i^T B_i d_i}{y_i^T d_i} - \right. \\ & \left. -2\beta_i \frac{y_i^T B_i d_i}{y_i^T d_i} - (1-\beta_i)\frac{(B_i d_i)^T B_i d_i}{d_i^T B_i d_i} \right) \le \\ &\le Tr(B_i) + \frac{\overline{\gamma}}{\underline{\rho}}\overline{G} + \frac{\overline{G}}{1-\varepsilon_2}\alpha_i + \frac{2\overline{G}}{1-\varepsilon_2}\frac{\alpha_i}{\kappa_i} - \frac{\lambda \underline{G}}{2(1-\varepsilon_1)}\frac{\alpha_i}{\kappa_i^2} \end{aligned} \tag{T}$$

and

$$\det(B_{i+1}) = \frac{1}{\gamma_i^n}\frac{\gamma_i}{\rho_i}\det(B_i)\frac{y_i^T d_i}{d_i^T B d_i}\left(1 - \frac{\beta_i}{\beta_i^*}\right) \ge \frac{1}{\overline{\gamma}^n}\frac{1}{\overline{\rho}}\det(B_i)\lambda\frac{1-\varepsilon_2}{\alpha_i} \tag{D},$$

where $\alpha_i > 0$ is a stepsize and $\kappa_i > 0$, is the cosine ($|g_i^T d_i| = \kappa_i \parallel g_i \parallel\parallel d_i \parallel$).

Proof The expression (B) can be written, after some manipulations, in the form

$$B_{i+1} = \frac{1}{\gamma_i}\left(B_i + \frac{\gamma_i}{\rho_i}\frac{y_i y_i^T}{y_i^T d_i} + \beta_i \frac{d_i^T B_i d_i}{y_i^T d_i}\frac{y_i y_i^T}{y_i^T d_i} - \frac{\beta_i}{y_i^T d_i}(B_i d_i y_i^T + y_i(B_i d_i)^T)) + \frac{\beta_i - 1}{d_i^T B_i d_i} B_i d_i (B_i d_i)^T \right).$$

Utilizing the fact that the trace is a linear matrix function and that $Tr(uv^T) = v^T u$ holds for arbitrary vectors $u \in R^n$, $v \in R^n$, we obtain the first part of the relation (T). The second part can be derived from the mean value theorems

$$y_i = g_{i+1} - g_i = \int_0^1 G(x_i + \lambda d_i) d_i d\lambda \triangleq \tilde{G}_i d_i, \tag{a}$$

and

$$F_{i+1} - F_i \ge g_i^T d_i + \frac{1}{2}\underline{G} \parallel d_i \parallel^2 \tag{b}$$

and using the formulas $F_{i+1} - F_i \le \varepsilon_1 g_i^T d_i$ (S3a) and $y_i^T d_i \ge (1-\varepsilon_2)|g_i^T d_i|$ (S3b) and $B_i d_i = -\alpha_i g_i$ (VM1). Formulas (a), (S3b) and (VM1) imply

$$\begin{aligned} \frac{y_i^T y_i}{y_i^T d_i} &= \frac{d_i^T \tilde{G}_i^2 d_i}{d_i^T \tilde{G}_i d_i} \le \overline{G}, \\ \frac{d_i^T B_i d_i}{y_i^T d_i} &= -\frac{\alpha_i g_i^T d_i}{(g_{i+1} - g_i)^T d_i} \le \frac{\alpha_i}{(1-\varepsilon_2)}, \\ \frac{|y_i^T B_i d_i|}{y_i^T d_i} &\le \frac{\parallel \tilde{G} d_i \parallel\parallel \alpha_i g_i \parallel}{(1-\varepsilon_2)|g_i^T d_i|} \le \frac{\alpha_i \overline{G}}{(1-\varepsilon_2)\kappa_i}. \end{aligned}$$

Using (b) and (S3a) we obtain

$$(1-\varepsilon_1)|g_i^T d_i| \ge \frac{1}{2}\underline{G} \parallel d_i \parallel^2$$

or $\parallel g_i \parallel / \parallel d_i \parallel \ge \underline{G}/(2\kappa_i(1-\varepsilon_1))$, so that we can write

$$\frac{(B_i d_i)^T B_i d_i}{d_i^T B_i d_i} = \frac{\alpha_i \parallel g_i \parallel}{\parallel d_i \parallel \kappa_i} \ge \frac{\alpha_i \underline{G}}{2(1-\varepsilon_1)\kappa_i^2}.$$

Substituting the above inequalities into the first part of the formula (T) and using the fact that $\beta_i \leq (1-\lambda) < 1$, we obtain the second part of (T). In the proof of the formula (D) we start from the fact that

$$\det(\gamma_i B_{i+1}) = \det B_i / q_i$$

(Lemma 5). The expression for $1/q_i$ can be derived from duality (Remark 28) or using the relation between β_i and η_i (Theorem 17). We obtain

$$\frac{1}{q_i} = \frac{\gamma_i}{\rho_i} \frac{y_i^T d_i}{d_i^T B_i d_i} \left(1 - \frac{\beta_i}{\beta_i^*}\right)$$

(Remark 29), which after a substitution into the expression of $\det(\gamma_i B_{i+1})$ gives the first part of formula (D). The second part of this formula follows from the inequalities already proved.

Theorem 20 (Global convergence) Let the assumptions of Lemma 6 be satisfied. Then

$$\liminf_{i \to \infty} \| g_i \| = 0.$$

Proof Using formula (T), we obtain

$$\begin{aligned} Tr(B_{i+1}) &\leq Tr(B_i) + \frac{\overline{\gamma}}{\underline{\rho}} \overline{G} + \overline{G} \frac{\| d_i \|^2}{y_i^T d_i} \frac{\| B_i d_i \|}{\| d_i \|} + 2 \frac{\| y_i \| \| d_i \|}{y_i^T d_i} \frac{\| B_i d_i \|}{\| d_i \|} \leq \\ &\leq Tr(B_i) + \frac{\overline{\gamma}}{\underline{\rho}} \overline{G} + \left(\frac{\overline{G}}{\underline{G}} + 2 \frac{\overline{G}}{\underline{G}} \right) \| B_i \| \leq \frac{\overline{\gamma}}{\underline{\rho}} \overline{G} + \left(1 + 3 \frac{\overline{G}}{\underline{G}}\right) Tr(B_i) \end{aligned}$$

(we use the main value theorem). This linear recurrence relation implies the existence of a constant $\overline{C} > 1$, such that

$$Tr(B_{i+1}) \leq \overline{C}^{i+1-k}$$

(The value $Tr(B_k)$, $k \leq i$, is considered to be initial). The second part of formula (D) implies

$$\det(B_{i+1}) \geq \frac{\lambda(1-\varepsilon_2)}{\rho \overline{\gamma}^n} \det(B_i) \frac{1}{\alpha_i} \geq \left(\frac{\lambda(1-\varepsilon_2)}{\rho \overline{\gamma}^n} \right)^{i+1-k} \det(B_k) \prod_{j=k}^{i} \frac{1}{\alpha_j}.$$

Using the inequality between the geometric and the arithmetic means, we obtain

$$\det(B_{i+1}) \leq \left(\frac{Tr(B_{i+1})}{n} \right)^n \leq (Tr(B_{i+1}))^n \leq (\overline{C}^n)^{i+1-k},$$

so that

$$\prod_{j=k}^{i} \frac{1}{\alpha_j} \leq \left(\frac{\overline{\rho \gamma}^n \overline{C}^n}{\lambda(1-\varepsilon_2) \min(1, \det(B_k))} \right)^{i+1-k} \triangleq \frac{1}{\underline{C}^{i+1-k}}$$

or, using the inequality between the geometric and the arithmetic means

$$\underline{C} \leq \left(\prod_{j=k}^{i} \alpha_j \right)^{\frac{1}{i+1-k}} \leq \frac{1}{i+1-k} \sum_{j=k}^{i} \alpha_j \qquad (*).$$

Denote

$$\xi_i = \frac{\overline{G}}{1-\varepsilon_2} + \frac{2\overline{G}}{(1-\varepsilon_2)\kappa_i} - \frac{\lambda \underline{G}}{2(1-\varepsilon_1)\kappa_i^2}$$

and suppose that $\liminf_{i\to\infty} \| g_i \| > 0$. Then, by the proof of Theorem 5, $\kappa_i \to 0$ necessarily holds and hence $\xi_i \to -\infty$. Therefore an index $k \in N$ exists, such that $\xi_i < -2\overline{\gamma}\overline{G}/\underline{\rho} < 0 \; \forall i \geq k$. But then (T) and $(*)$ imply

$$\begin{aligned} Tr(B_{i+1}) &\leq Tr(B_i) + \frac{\overline{\gamma}}{\underline{\rho}}\overline{G} + \alpha_i \xi_i \leq \\ &\leq Tr(B_k) + \frac{\overline{\gamma}}{\underline{\rho}}\overline{G}(i+1-k) - 2\frac{\overline{\gamma}}{\underline{\rho}}\overline{G}\sum_{j=k}^{i} \alpha_j \leq \\ &\leq Tr(B_k) - \frac{\overline{\gamma}}{\underline{\rho}}\overline{G}(i+1-k). \end{aligned}$$

If the index i is chosen in such a way that

$$i \geq k - 1 + \frac{\underline{\rho} Tr(B_k)}{\overline{\gamma}\overline{G}} \geq k,$$

then the last relation is negative which is a contradiction since trace of a SPD matrix is positive.

Remark 33 Theorem 20 is the most general known result concerning global convergence of variable metric methods. This theorem requires satisfying the condition (F4) (existence of a constant $\underline{C} > 0$), so that it can be used only for strongly convex functions. The condition $\gamma_i \leq 1$, $i \in N$, is important for the proof of Theorem 20 and it is used in Algorithm 3.

Remark 34 Theorem 20 gives a theoretical reason for bad convergence properties of the DFP method DFP (with imperfect line search). The DFP method corresponds to the choice $\beta = 1$, $i \in N$, so that the assumption $\beta_i \leq 1 - \lambda < 1$, $i \in N$ is not satisfied. In the relation (T), the last term is not present and, therefore, the principle of the proof cannot be used.

Now we introduce a ressult concerning superlinear convergence of variable metric methods. Since the superlinear convergence requires the assumption that, in some sense, $B_i \to G^*$ (Theorem 8), we will require satisfying the assumptions of Theorem 14 ($\rho_i = 1$, $\gamma_i = 1$, $\forall i \in N$).

Proposition 5 (Superlinear convergence) Let the assumptions of Lemma 6 be satisfied with $\rho_i = 1$, $\gamma_i = 1$, and let in addition $x_i \to x^*$, the function $F : R^n \to R$ satisfy the condition (F5) and the value $\alpha_i = 1$ be always chosen, if this value satisfies the weak Wolfe condition. If

$$\sum_{\substack{i=1 \\ \beta_i < 0}}^{\infty} \frac{\beta_i}{\beta_i^*} < \infty,$$

then

$$\lim_{i\to\infty} \frac{\| x_{i+1} - x^* \|}{\| x_i - x^* \|} = 0$$

Remark 35 In the proof of Proposition 5, the invariance of the Broyden class of variable metric methods to transformation of variables is used. If we denote $x = T\hat{x}$, then $\hat{g} = T^T g$, $\hat{G} = T^T G T$ and also $d = T\hat{d}$, $\hat{y} = T^T y$. If we denote $\hat{B} = T^T B T$, $\hat{B}^+ = T^T B^+ T$, then (B) implies

$$\hat{B}^+ = \frac{1}{\gamma}\left(\hat{B} + \frac{\gamma}{\rho}\frac{1}{\hat{b}}\hat{y}\hat{y}^T - \frac{1}{\hat{c}}\hat{B}\hat{d}(\hat{B}\hat{d})^T + \frac{\beta}{\hat{c}}\left(\frac{\hat{c}}{\hat{b}}\hat{y} - \hat{B}\hat{d}\right)\left(\frac{\hat{c}}{\hat{b}}\hat{y} - \hat{B}\hat{d}\right)^T\right) \qquad (\hat{\text{B}}),$$

where $\hat{a} = \hat{y}^T \hat{B}^{-1}\hat{y}$, $\hat{b} = \hat{y}^T \hat{d}$, $\hat{c} = \hat{d}^T \hat{B}\hat{d}$ and where the parameters β, γ, ρ are the same as in (B). For theoretical purposes, $T = (G^*)^{-\frac{1}{2}}$ is usually used.

Now we introduce several comments to the implementation of variable metric methods.

1) Stepsize selection: Variable metric methods are not sensitive on the stepsize selection. Algorithm 1 can be used without changes. The following initial stepsize is often chosen

$$\alpha = \min\left(1, \frac{4(\underline{F} - F_i)}{s_i^T g_i}\right).$$

2) Nonquadratic correction (parametr ρ): Denote $\varphi(\lambda) = F(x+\lambda d)$. Parameter ρ is selected in such a way that $d^T B_+ d \approx \varphi''(1)$. Using the backward Taylor expansion $\varphi(0) = \varphi(1) - \varphi'(1) + \varphi''(\hat{\lambda})/2$, where $0 \leq \hat{\lambda} \leq 1$, we can write $\varphi''(\hat{\lambda}) = 2(\varphi(0) - \varphi(1) + \varphi'(1))$, so that, after substituting $d^T B_+ d = \varphi''(\hat{\lambda}) \approx \varphi''(1)$ into (VM3), we obtain $d^T y/\rho = d^T B_+ d = 2(\varphi(0) - \varphi(1) + \varphi'(1))$, which gives

$$\rho = \frac{d^T y}{2(F - F_+ + d^T g_+)}$$

This value is used only if $0.01 \leq \rho \leq 100$, otherwise we set $\rho = 1$.

3) Scaling (parameter γ): Since (VM3) implies $B_+ d = y/\rho$, it is advantageous to choose the parameter γ so that Bd/γ be as close to y/ρ as possible, e.g.

$$d^T Bd/\gamma = d^T y/\rho \Rightarrow \gamma/\rho = c/b$$

or

$$y^T d/\gamma = y^T B^{-1} y/\rho \Rightarrow \gamma/\rho = b/a.$$

An additional possibility is the geometrical mean $\gamma/\rho = (c/a)^{\frac{1}{2}}$. Thus for a given ρ we find $\gamma = \rho c/b$ or $\gamma = \rho b/a$ or $\gamma = \rho(c/a)^{\frac{1}{2}}$. If for such an obtained value $1 \leq \gamma \leq 10$ does not hold, we set $\gamma = 1$.

4) Choice of the particular method (parameter β): Practical experience shows that the BFGS method is the most efficient one of all simple updates and that the DFP method is bad. Although the BFGS method can be outperformed by some more complicated methods, suitable scaling and possible nonquadratic correction decrease differences among individual methods with the total improvement of the efficiency, so that we can recommend the scaled BFGS method realized by the following algorithm.

Algorithm 3 (VM) Data $\varepsilon_1 = 10^{-4}$, $\varepsilon_2 = 0.9$, $\underline{\varepsilon} > 0$, $\underline{\rho} = 0.01$, $\overline{\rho} = 100$, $\underline{\gamma} = 1$, $\overline{\gamma} = 10$.

Step 1 Choose an initial point $x_1 \in R^n$ and compute $F_1 = F(x_1)$, $g_1 = g(x_1)$, choose an initial SPD matrix H_1 (usually $H_1 = I$) and set $i = 1$.

Step 2 If $\| g_i \| \leq \underline{\varepsilon}$, then terminate the computation. Otherwise set $s_i = -H_i g_i$ and determine the stepsize α_i by Algorithm 1. Set $x_{i+1} = x_i + \alpha_i s_i$ and compute $F_{i+1} = F(x_{i+1})$, $g_{i+1} = g(x_{i+1})$.

Step 3 Set $d_i = x_{i+1} - x_i$ and $y_i = g_{i+1} - g_i$. Determine the parameter ρ_i by 2. If $\rho_i < \underline{\rho}$ or $\rho_i > \overline{\rho}$, then set $\rho_i = 1$. Determine the parameter γ_i by 3. If $i > 1$ and simultaneously $\gamma_i < \underline{\gamma}$ or $\gamma_i > \overline{\gamma}$, then set $\gamma_i = 1$. Choose $\eta_i = 1$ (the BFGS method). Determine the matrix H_{i+1} by the formula (H), increase i by 1 and go to Step 2.

The following table shows a comparison of three methods (CG - conjugate gradient method, VM - variable metric method, MN-modified Newton method) for minimization of 24 test functions with 10, 20, 40 and 80 variables (the total numbers of iterations NIT and function evaluations NFV are presented, together with the total computational time).

Method	n	NIT - NFV	time
CG	10	1211 - 2643	1.54
	20	1346 - 2897	3.52
	40	2105 - 4397	11.59
	80	4035 - 8125	54.37
VM	10	959 - 1125	1.10
	20	1115 - 1261	2.53
	40	1491 - 1666	7.85
	80	2200 - 2433	35.37
MN	10	425 - 527	1.15
	20	407 - 508	3.30
	40	433 - 571	15.54
	80	514 - 686	1:54.24

3 Trust region methods

3.1 Basic properties of the trust region methods

In the description of the trust region methods, we will use the notation

$$Q_i(s) = g_i^T s + \frac{1}{2} s^T B_i s$$

for the quadratic function which locally approximates the difference $F(x_i+s)-F(x_i)$ and the notation

$$\omega_i(s) = (B_i s + g_i)/ \parallel g_i \parallel$$

for the precision of direction determination. Furthermore, we will use the notation

$$\rho_i(s) = \frac{F(x_i + s) - F(x_i)}{Q_i(s)}$$

for the ratio of the actual and the predicted decrease of the function $F : R^n \to R$.

Definition 22 We say that the basic optimization method $x_{i+1} = x_i + \alpha_i s_i$, $i \in N$, is a trust region method, if the direction vectors $s_i \in R^n$, $i \in N$, are determined so that

$$\parallel s_i \parallel \leq \Delta_i, \tag{T1a}$$

$$\parallel s_i \parallel < \Delta_i \Rightarrow \parallel \omega_i(s_i) \parallel \leq \overline{\omega}_i \leq \overline{\omega}, \tag{T1b}$$

$$-Q_i(s_i) \geq \underline{\sigma} \parallel g_i \parallel \min(\parallel s_i \parallel, \parallel g_i \parallel / \parallel B_i \parallel), \tag{T1c}$$

where $0 < \underline{\sigma} < 1$ and $0 \leq \overline{\omega} < 1$, and the stepsizes $\alpha_i \geq 0$, $i \in N$, are determined so that

$$\rho_i(s_i) \leq 0 \Rightarrow \alpha_i = 0, \tag{T2a}$$

$$\rho_i(s_i) > 0 \Rightarrow \alpha_i = 1 \qquad \text{(T2b).}$$

Trust region radii $\Delta_i > 0$, $i \in N$, are constructed in such a way that $\Delta_1 > 0$ is properly chosen and

$$\rho_i(s_i) < \underline{\rho} \Rightarrow \underline{\beta} \parallel s_i \parallel \leq \Delta_{i+1} \leq \overline{\beta} \parallel s_i \parallel, \tag{T3a}$$

$$\rho_i(s_i) \geq \underline{\rho} \Rightarrow \Delta_i \leq \Delta_{i+1} \leq \overline{\gamma} \Delta_i, \tag{T3b}$$

where $0 < \underline{\beta} \leq \overline{\beta} < 1 < \overline{\gamma}$ and $0 < \underline{\rho} < 1$.

Remark 36 If $\overline{\omega} = 0$ or $\overline{\omega} > 0$ we obtain exact or inexact trust region methods, respectively.

Remark 37 Norms in (T1) and (T3) can be more general than Euclidean. Some conditions can be weakened.

Remark 38 We denote $N_1 \subset N$ the set of indices such that $\| s_i \| < \Delta_i$, $N_2 \subset N$ the set of indices such that $\rho_i(s_i) > 0$ and $N_3 \subset N$ the set of indices such that $\rho_i(s_i) \geq \underline{\rho}$.

Lemma 7 If we apply the trust region method (T1)-(T3) to the function $F : R^n \to R$ satisfying the condition (F3), then constant $\underline{c} > 0$ exists, such that

$$\| s_i \| \geq \underline{c} m_i / M_i, \tag{$*$}$$

where

$$m_i = \min_{1 \leq j \leq i} \| g_j \|,$$

$$M_i = \max_{1 \leq j \leq i} \| B_j \| .$$

Proof (a) Let $i \in N_1$. Then (T1b) implies

$$| \| B_i s_i \| - \| g_i \| | \leq \| B_i s_i + g_i \| = \| \omega_i(s_i) \| \| g_i \| \leq \overline{\omega} \| g_i \|,$$

so that either $\| B_i s_i \| \geq \| g_i \|$ or $\| B_i s_i \| < \| g_i \|$ and $\| B_i s_i \| \geq (1 - \overline{\omega}) \| g_i \|$. Connecting these inequalities, we obtain $\| B_i \| \| s_i \| \geq \| B_i s_i \| \geq (1 - \overline{\omega}) \| g_i \|$, which gives $\| s_i \| \geq (1 - \overline{\omega}) m_i / M_i$.
(b) let $i \notin N_3$. Then from the definition of the set N_3 and the function $Q_i(s)$ it follows

$$F(x_i + s_i) - F(x_i) \geq \underline{\rho} Q_i(s_i) = \underline{\rho} \left(g_i^T s_i + \frac{1}{2} s_i^T B_i s_i \right) \geq \underline{\rho} \left(g_i^T s_i - \frac{1}{2} \| B_i \| \| s_i \|^2 \right).$$

On the other hand, the mean value theorem implies

$$F(x_i + s_i) - F(x_i) \leq g_i^T s_i + \frac{1}{2} \overline{G} \| s_i \|^2,$$

which together with the previous inequality gives

$$\frac{1}{2} (\overline{G} + \underline{\rho} \| B_i \|) \| s_i \|^2 \geq (\underline{\rho} - 1) g_i^T s_i$$

From (T1c) we obtain

$$-\underline{\sigma} \| g_i \| \min(\| s_i \|, \| g_i \| / \| B_i \|) \geq Q_i(s_i) \geq g_i^T s_i - \frac{1}{2} \| B_i \| \| s_i \|^2,$$

which together with the previous inequality gives

$$\begin{aligned} \frac{1}{2} (\overline{G} + \underline{\rho} \| B_i \|) \| s_i \|^2 &\geq (\underline{\rho} - 1) g_i^T s_i \geq \frac{1}{2} (\underline{\rho} - 1) \| B_i \| \| s_i \|^2 - \\ &- \underline{\sigma} (\underline{\rho} - 1) \| g_i \| \min(\| s_i \|, \| g_i \| / \| B_i \|) \end{aligned}$$

or

$$\frac{1}{2} (\overline{G} + \| B_i \|) \| s_i \|^2 \geq \underline{\sigma} (1 - \underline{\rho}) \| g_i \| \min(\| s_i \|, \| g_i \| / \| B_i \|),$$

so that either $\| s_i \| \geq \| g_i \| / \| B_i \| \geq m_i / M_i$ or

$$\frac{1}{2} (\overline{G} M_i / \| B_1 \| + M_i) \| s_i \| \geq \underline{\sigma} (1 - \underline{\rho}) \| g_i \|,$$

which gives $\| s_i \| \geq [2 \underline{\sigma} (1 - \underline{\rho}) \| B_1 \| / (\overline{G} + \| B_1 \|)] m_i / M_i$.
(c) Let $i = 1$. If $\| g_1 \| = 0$, then obviously $\| s_1 \| \geq \| g_1 \| / \| B_1 \| \geq m_1 / M_1$. If $\| g_1 \| \neq 0$, then we can write

$$\| s_1 \| = \frac{\| s_1 \| \| B_1 \|}{\| g_1 \|} \frac{\| g_1 \|}{\| B_1 \|},$$

so that $\| s_1 \| \geq (\| s_1 \| \| B_1 \| / \| g_1 \|) m_1 / M_1$.
(d) Let $i \notin N_1$, $i \in N_3$ and $i \neq 1$. Let $k < i$ be the greatest index for which simultaneously $k \notin N_1$, $k \in N_3$ and $k \neq 1$ does not hold. Then from (T3) and (T1a) we get

$$\| s_i \| \geq \Delta_i \geq \Delta_{k+1} \geq \min(\Delta_k, \underline{\beta} \| s_k \|) \geq \min(\| s_k \|, \underline{\beta} \| s_k \|) \geq \underline{\beta} \| s_k \|,$$

so that (a)-(c) imply

$$\| s_i \| \geq \underline{\beta} \| s_k \| \geq \underline{c} m_k / M_k \geq \underline{c} m_i / M_i,$$

where

$$\underline{c} = \underline{\beta} \min \left((1 - \overline{\omega}), \frac{2\underline{\sigma}(1 - \underline{\rho}) \| B_1 \|}{\overline{G} + \| B_1 \|}, \frac{\| s_1 \| \| B_1 \|}{\| g_1 \|} \right).$$

Theorem 21 Let $x_i \in R^n$, $i \in N$, be a sequence generated by the trust region method (T1)-(T3) such that

$$\sum_{i=1}^{\infty} \frac{1}{M_i} = \infty$$

where M_i, $i \in N$, are the numbers defined in Lemma 7. Let the function $F : R^n \to R$ satisfy the conditions (F1) and (F3). Then

$$\liminf_{i \to \infty} \| g_i \| = 0$$

Proof (a) Suppose that a number $\underline{\varepsilon} > 0$ exists, so that $\| g_i \| \geq \underline{\varepsilon}$ $\forall i \in N$. Then from Lemma 7 it holds

$$\| s_i \| \geq \frac{\underline{c}\underline{\varepsilon}}{M_i} \tag{$*$}$$

$\forall i \in N$. Since $N_3 \subset N_2$, we can write

$$F_i - F_{i+1} = F(x_i) - F(x_i + s_i) \geq -\underline{\rho} Q_i(s_i) \geq \underline{\rho}\underline{\sigma}\underline{\varepsilon} \min (\| s_i \|, \underline{\varepsilon}/M_i) \geq \underline{\rho}\underline{\sigma}\underline{\varepsilon}^2 \underline{c}/M_i$$

$\forall i \in N_3$, so that

$$F_1 - \underline{F} \geq \lim_{i \to \infty} (F_1 - F_{i+1}) = \sum_{i=1}^{\infty} (F_i - F_{i+1}) \geq \sum_{i \in N_3} (F_i - F_{i+1}) \geq \underline{\rho}\underline{\sigma}\underline{\varepsilon}^2 \underline{c} \sum_{i \in N_3} \frac{1}{M_i}.$$

Therefore,

$$\sum_{i \in N_3} \frac{1}{M_i} < \infty.$$

(b) Let $i \in N$ and let r be an integer, such that $\overline{\beta}^{r-1} \overline{\gamma} < 1$ (Such a number exists since $\overline{\beta} < 1$ and $\overline{\gamma} < \infty$). Let $p(i)$ be a number of indices from the set $[1, i] = \{1, \ldots, i\}$, which are elements of the set N_3 (or $p(i)$ is the cardinality of the set $[1, i] \cap N_3$). Denote

$$N_4 = \{ i \in N : rp(i) < i \}.$$

Then (T1a) and (T3) imply

$$\Delta_i \leq \overline{\gamma}^{p(i-1)} \overline{\beta}^{i-1-p(i-1)} \Delta_1 \leq \overline{\gamma}^{(i-1)/r} \overline{\beta}^{(r-1)(i-1)/r} \Delta_1 \leq \left(\overline{\gamma}\overline{\beta}^{(r-1)} \right)^{(i-1)/r} \Delta_1.$$

Since we suppose that $\overline{\gamma}\overline{\beta}^{r-1} < 1$, we can write

$$\sum_{i \in N_4} \Delta_i \le \sum_{i \in N_4} \left(\overline{\gamma}\overline{\beta}^{(r-1)}\right)^{(i-1)/r} \Delta_1 \le \sum_{i=1}^{\infty} \left(\overline{\gamma}\overline{\beta}^{(r-1)}\right)^{(i-1)/r} \Delta_1 = \frac{\Delta_1}{1 - \left(\overline{\gamma}\overline{\beta}^{(r-1)}\right)^{1/r}} < \infty.$$

If we now use (T1a) and $(*)$, we obtain

$$\sum_{i \in N_4} \frac{1}{M_i} \le \frac{1}{\underline{c}\varepsilon} \sum_{i \in N_4} \| s_i \| \le \frac{1}{\underline{c}\varepsilon} \sum_{i \in N_4} \Delta_i < \infty.$$

(c) Now it suffices to prove that

$$\sum_{i \in N_5} \frac{1}{M_i} < \infty,$$

where $N_5 = N \setminus N_4$, so that $N_5 = \{i \in N : rp(i) \ge i\}$. Denote

$$\begin{aligned} N_3 &= \{i_1, i_2, i_3 \ldots\} \\ N_5 &= \{k_1, k_2, k_3 \ldots\} \end{aligned}$$

(we suppose sorting of elements by their sizes) and construct the set

$$N_6 = \{l_1, l_2, l_3 \ldots\} = \{\underbrace{i_1, \ldots, i_1}_{r-\text{times}}, \underbrace{i_2, \ldots, i_2}_{r-\text{times}}, \underbrace{i_3, \ldots, i_3}_{r-\text{times}}, \ldots\}.$$

From the construction of the set N_5 it follows that

$$rp(k_j) \ge k_j \ge j \quad \forall j \in N,$$

so that, using the definition of the set N_6, we obtain

$$l_j \le l_{rp(k_j)} \le i_{p(k_j)} \le k_j \quad \forall j \in N,$$

since $i_{p(k_j)}$ is the last element of the set $[1, k_j] \cap N_3$. Therefore, $M_{l_j} \le M_{k_j}$ $\forall j \in N$, so that (a) implies

$$\sum_{i \in N_5} \frac{1}{M_i} = \sum_{j=1}^{\infty} \frac{1}{M_{k_j}} \le \sum_{j=1}^{\infty} \frac{1}{M_{l_j}} = \sum_{i \in N_6} \frac{1}{M_i} = r \sum_{i \in N_3} \frac{1}{M_i} < \infty.$$

Remark 39 The assumption of the Theorem 21 are satisfied e.g. if $\| B_i \| \le \overline{B}$ $\forall i \in N$. Proof of this partial statement is very easy. It suffices to modify part (a) in the proof of Theorem 21 in such a way that

$$F_1 - \underline{F} \ge \lim_{i \to \infty} (F_1 - F_{i+1}) = \sum_{i=1}^{\infty} (F_i - F_{i+1}) \ge \sum_{i \in N_3} (F_i - F_{i+1}) \ge \underline{\rho}\underline{\sigma}\varepsilon^2 \underline{c} \sum_{i \in N_3} \frac{1}{\overline{B}}.$$

If the set N_3 is infinite, then we come immediately to a contradiction. If the set N_3 is finite, then (T3a) implies $\Delta_i \to 0$, which together with (T1a) gives $\| s_i \| \to 0$, which is a contradiction to $(*)$.

Theorem 22 (superlinear convergence). Let $x_i \in R^n$, $i \in N$, be a sequence generated by the trust region method (T1)-(T3), such that $x_i \to x^*$. Let the function $F : R^n \to R$ satisfy the conditions (F3) and (F4). Let

$$\lim_{i \to \infty} \overline{\omega}_i = 0, \tag{α}$$

$$\lim_{i \to \infty} \frac{\| (B_i - G_i) s_i \|}{\| s_i \|} = 0 \tag{β}$$

and $\| B_i \| \le \overline{G}$ $\forall i \in N$, where $\overline{G} > \overline{\lambda}(G^*)$. Then the sequence x_i, $i \in N$, converges Q-superlinearly to the point $x^* \in R^n$.

Proof (a) We show that an index $k_2 \in N$ exists, such that

$$\| g_i \| \geq \frac{1}{2}\underline{G} \| s_i \|$$

and

$$-Q_i(s_i) \geq \frac{\sigma \underline{G}^2}{4\overline{G}} \| s_i \|^2$$

$\forall i \geq k_2$, if $\underline{G} < \underline{\lambda}(G^*)$. Denote $\vartheta_i = (B_i - G_i)s_i / \| s_i \|$. Then

$$s_i^T B_i s_i = s_i^T G_i s_i + s_i^T \vartheta_i \| s_i \| \geq s_i^T G_i s_i - \| \vartheta_i \| \| s_i \|^2$$

and since $\| \vartheta_i \| \to 0$, $G_i \to G^*$ and $\underline{G} < \underline{\lambda}(G^*)$, an index $k_2 \in N$ exists, such that $s_i^T B_i s_i \geq \underline{G} \| s_i \|^2$ $\forall i \geq k_2$. From the definition of $Q_i(s_i)$ and from (T1c) it follows that

$$0 \geq Q_i(s_i) = g_i^T s_i + \frac{1}{2} s_i^T B_i s_i \geq \frac{1}{2}\underline{G} \| s_i \|^2 - \| g_i \| \| s_i \|,$$

which gives $\| g_i \| \geq (\underline{G}/2) \| s_i \|$ $\forall i \geq k_2$. Using once more (T1c), we can write

$$-Q_i(s_i) \geq \underline{\sigma} \| g_i \| \min(\| s_i \|, \| g_i \| / \| B_i \|) \geq \frac{\sigma \underline{G}}{2} \min(1, \frac{\underline{G}}{2\overline{G}}) \| s_i \|^2 = \frac{\sigma \underline{G}^2}{4\overline{G}} \| s_i \|^2 .$$

(b) We show that an index $k_3 \geq k_2$ exists, such that $i \in N_3$ $\forall i \geq k_3$. Using the mean value theorem, we obtain

$$F(x_i + s_i) - F(x_i) = s_i^T g_i + \frac{1}{2} s_i^T G_i s_i + o(\| s_i \|^2) = Q_i(s_i) + \frac{1}{2} s_i^T (G_i - B_i) s_i + o(\| s_i \|^2),$$

so that

$$\rho_i(s_i) = \frac{F(x_i + s_i) - F(x_i)}{Q_i(s_i)} = 1 + \frac{s_i^T (G_i - B_i) s_i + o(\| s_i \|^2)}{2 Q_i(s_i)}.$$

However, (a) implies

$$\left| \frac{s_i^T (G_i - B_i) s_i + o(\| s_i \|^2)}{2 Q_i(s_i)} \right| \leq \frac{2\overline{G}}{\underline{\sigma G}^2} \frac{\| \vartheta_i \| \| s_i \|^2 + o(\| s_i \|^2)}{\| s_i \|^2} \to 0,$$

since $\| \vartheta_i \| \to 0$. Therefore, $\rho_i(s_i) \to 1$ and, since $\underline{\rho} < 1$, an index $k_3 \geq k_2$ exists, such that $\rho_i(s_i) \geq \underline{\rho}$ $\forall i \geq k_3$.
(c) We show, that an index $k \geq k_3$ exists, such that $i \in N_1$ $\forall i \geq k$. Notice first that the set $N_1 \subset N$ is infinite. Otherwise, since (b) implies $i \in N_3$ $\forall i \geq k_3$, the inequality $\| s_i \| \geq \Delta_i \geq \Delta_{k_3}$ $\forall i \geq k_3$, has to hold. However, it is the contradiction, since (a) implies $\| s_i \| \leq \| g_i \| / \underline{G}$, so that $\| g_i \| \to 0$ implies $\| s_i \| \to 0$. Consider now only the indices $i \geq k_3$, $i \in N_1$ and denote $\omega_i = \omega_i(s_i)$. Using (α), (β) and (T1b), we obtain $\| \omega_i \| \to 0$ and $\| \vartheta_i \| \to 0$. Hence we can show, in the similar way as in the proof of Theorem 8, that an index $k_4 \geq k_3$, $k_4 \in N_1$, exists, such that

$$\| g_i \| / \overline{G} \leq \| s_i \| \leq \| g_i \| / \underline{G}$$

$\forall i \geq k_4$, $i \in N_1$. Using the mean value theorem, we can write

$$g_{i+1} = g(x_i + s_i) = g_i + G_i s_i + o(\| s_i \|),$$

since $i \in N_3 \subset N_2$. Denote

$$\lambda_i = \frac{g_{i+1} - g_i - B_i s_i}{\| g_i \|}.$$

Then it follows from the above consideration that $\| \lambda_i \| \le \| \vartheta_i \| / \underline{G} + o(\| s_i \|) / \| s_i \| \to 0$. Since simultaneously $\| \omega_i \| \to 0$ holds, an index $k \ge k_4$, $k \in N_1$ exists, such that $\| \lambda_i \| < (\underline{G}/\overline{G})/2$ and $\| \omega_i \| < (\underline{G}/\overline{G})/2$ $\forall i \ge k$, $i \in N_1$. Then we can write

$$\begin{aligned} \| s_{i+1} \| &\le \frac{1}{\underline{G}} \| g_{i+1} \| \le \frac{1}{\underline{G}} (\| g_{i+1} - g_i - B_i s_i \| + \| B_i s_i + g_i \|) \le \\ &\le \frac{\overline{G}}{\underline{G}} (\| \lambda_i \| + \| \omega_i \|) \| s_i \| < \left(\frac{1}{2} + \frac{1}{2} \right) \| s_i \| = \| s_i \| . \end{aligned}$$

Since $i \in N_3$ by (b), $\Delta_{i+1} \ge \Delta_i$ holds, which gives $\| s_{i+1} \| < \| s_i \| \le \Delta_i \le \Delta_{i+1}$, so that $i+1 \in N_1$. Using induction, we obtain $i \in N_1$ $\forall i \ge k$.
(d) Superlinear convergence. It holds

$$\frac{\| g_{i+1} \|}{\| g_i \|} \le \frac{\| g_{i+1} - g_i - B_i s_i \| + \| B_i s_i + g_i \|}{\| g_i \|} \le \| \lambda_i \| + \| \omega_i \|,$$

which together with $\| \lambda_i \| \to 0$ and $\| \omega_i \| \to 0$ gives

$$\lim_{i \to \infty} \frac{\| x_{i+1} - x^* \|}{\| x_i - x^* \|} \le \frac{\overline{G}}{\underline{G}} \lim_{i \to \infty} \frac{\| g_{i+1} \|}{\| g_i \|} = 0.$$

3.2 Trust region methods with optimal locally constrained step

Definition 23 Trust region methods with the optimal locally constrained step use direction vectors $s_i \in R^n$, $i \in N$, such that

$$s_i = \arg \min_{\| s \| \le \Delta_i} Q_i(s) \qquad (\overline{\mathrm{T}}1)$$

where $\| s_i \| = \Delta_i$, if this minimun is not unique.

Theorem 23 The direction vector determined by $(\overline{\mathrm{T}}1)$ satisfies the conditions (T1) with $\overline{\omega} = 0$ and $\underline{\sigma} = 1/2$.

Proof Condition (T1a) is directly a part of the condition $(\overline{\mathrm{T}}1)$.
(a) Assume that $s_i \in R^n$ is a solution of the problem $(\overline{\mathrm{T}}1)$, such that $\| s_i \| < \Delta_i$. Then necessarily $Q_i(s)$ is a strict convex function and $B_i s_i + g_i = 0$, so that $\omega_i(s_i) = 0$ and

$$-Q_i(s_i) = g_i^T B_i^{-1} g_i - \frac{1}{2} g_i^T B_i^{-1} g_i = \frac{1}{2} g_i^T B_i^{-1} g_i \ge \frac{1}{2} \| g_i \|^2 / \| B_i \| .$$

(b) Let $\| s_i \| = \Delta_i$. Set

$$s = -\frac{g_i^T g_i}{g_i^T B_i g_i} g_i$$

and suppose that $\| s \| \le \Delta_i$ and $g_i^T B_i g_i > 0$. Then

$$-Q_i(s) = \frac{(g_i^T g_i)^2}{g_i^T B_i g_i} - \frac{1}{2} \frac{(g_i^T g_i)^2 g_i^T B_i g_i}{(g_i^T B_i g_i)^2} = \frac{1}{2} \frac{(g_i^T g_i)^2}{g_i^T B_i g_i} \ge \frac{1}{2} \| g_i \|^2 / \| B_i \| .$$

Condition $(\overline{\mathrm{T}}1)$ implies $Q_i(s_i) \le Q_i(s)$, so that necessarily

$$-Q_i(s_i) \ge -Q_i(s) \ge \frac{1}{2} \| g_i \|^2 / \| B_i \| .$$

(c) Let $\| s_i \| = \Delta_i$ and either $\| s \| > \Delta_i$ or $g_i^T B_i g_i \le 0$, where $s \in R^n$ is a vector defined in (b). If $\| s \| > \Delta_i$ and $g_i^T B_i g_i > 0$, then $\| g_i \|^3 / g_i^T B_i g_i > \Delta_i$ or

$$g_i^T B_i g_i < \| g_i \|^3 / \Delta_i .$$

The same inequality holds for $g_i^T B_i g_i < 0$. Set $\tilde{s} = -(\Delta_i / \| g_i \|) g_i$, so that $\| \tilde{s} \| \le \Delta_i$. Then

$$-Q_i(\tilde{s}) = \Delta_i \parallel g_i \parallel - \frac{1}{2} \frac{\Delta_i^2}{\parallel g_i \parallel^2} g_i^T B_i g_i > \Delta_i \parallel g_i \parallel - \frac{1}{2} \Delta_i \parallel g_i \parallel = \frac{1}{2} \Delta_i \parallel g_i \parallel = \frac{1}{2} \parallel s_i \parallel \parallel g_i \parallel$$

since $\parallel s_i \parallel = \Delta_i$. From ($\overline{\text{T}}1$) it follows that $Q_i(s_i) \leq Q_i(\tilde{s})$, so that necessarily

$$-Q_i(s_i) \geq -Q_i(\tilde{s}) \geq \frac{1}{2} \parallel g_i \parallel \parallel s_i \parallel .$$

3.3 Determination of the optimal locally constrained step

Theorem 24 Vector $s_i \in R^n$ is a solution of the problem ($\overline{\text{T}}1$) if and only if $\parallel s_i \parallel \leq \Delta_i$ and a number $\lambda_i \geq 0$ exists, such that the matrix $B_i + \lambda_i I$ is positive semidefinite and $(B_i + \lambda_i I)s_i + g_i = 0$ together with $(\parallel s_i \parallel - \Delta_i)\lambda_i = 0$ holds.

Proof First we prove necessity. If $\parallel s_i \parallel < \Delta_i$, then necessarily $B_i s_i + g_i = 0$ and $(\parallel s_i \parallel - \Delta_i) \neq 0$ and function $Q_i(s)$ is convex, so that the matrix B_i is positive semidefinite. Therefore, the required conditions with $\lambda_i = 0$ hold. If $\parallel s_i \parallel = \Delta_i$, then the Kuhn-Tucker conditions $(B_i + \lambda_i I)s_i + g_i = 0$ and $(\parallel s_i \parallel - \Delta_i)\lambda_i = 0$, with $\lambda_i \geq 0$, have to hold. It remains to prove positive semidefiniteness of the matrix $B_i + \lambda_i I$. For an arbitrary vector $s \in R^n$, such that $\parallel s \parallel = \Delta_i$, it holds

$$\begin{aligned}
Q_i(s) - Q_i(s_i) &= (s - s_i)^T g_i + \frac{1}{2} s^T B_i s - \frac{1}{2} s_i^T B_i s_i = \\
&= (s_i - s)^T (B_i + \lambda_i I) s_i + \frac{1}{2} s^T B_i s - \frac{1}{2} s_i^T B_i s_i = \\
&= \frac{1}{2} (s_i - s)^T (B_i + \lambda_i I)(s_i - s) + \frac{1}{2} \lambda_i (s_i^T s_i - s^T s) = \\
&= \frac{1}{2} (s_i - s)^T (B_i + \lambda_i I)(s_i - s) \geq 0,
\end{aligned}$$

so that the matrix $B_i + \lambda_i I$ has to be positive semidefinite. Now we prove sufficiency. If $\parallel s_i \parallel < \Delta_i$, then the function $Q_i(s)$ is convex (the matrix $B_i + \lambda_i I$ is positive semidefinite for $\lambda_i = 0$), so that the necessary conditions are also sufficient. If $\parallel s_i \parallel = \Delta_i$, then the required conditions imply (as above), that

$$\begin{aligned}
Q_i(s) - Q_i(s_i) &= (s - s_i)^T g_i + \frac{1}{2} s^T B_i s - \frac{1}{2} s_i^T B_i s_i = \\
&= \frac{1}{2} (s_i - s)^T (B_i + \lambda_i I)(s_i - s) + \frac{1}{2} \lambda_i (s_i^T s_i - s^T s) \geq \\
&\geq \frac{1}{2} (s_i - s)^T (B_i + \lambda_i I)(s_i - s) \geq 0
\end{aligned}$$

for all vectors $s \in R^n$ such that $\parallel s \parallel \leq \parallel s_i \parallel = \Delta_i$.

Statement of Theorem 24 is a basis for the algorithm, in which the nonlinear equation $\phi(\lambda) \stackrel{\Delta}{=} \frac{1}{\Delta_i} - \frac{1}{\|s_i(\lambda)\|}$ is solved by the Newton method (here vector $s_i(\lambda)$ is a solution of the linear system $(B_i + \lambda I)s_i(\lambda) + g_i = 0$):

Algorithm 4 Data $0 < \underline{\delta} < 1 < \overline{\delta}$ (usually $\underline{\delta} = 0.9$ and $\overline{\delta} = 1.1$).

Step 1 Determine $\underline{\gamma}$ as a maximum diagonal element of the matrix $-B$. Set $\underline{\lambda} = 0$ and $\overline{\lambda} = \parallel g \parallel / \Delta + \parallel B \parallel$. Set $\lambda = \max(\underline{\gamma}, \underline{\lambda})$.

Step 2 Set $\underline{\lambda} = \max(\underline{\gamma}, \underline{\lambda})$. If $\lambda < \underline{\gamma}$, then set $\lambda = \sqrt{\underline{\lambda}\overline{\lambda}}$.

Step 3 If the matrix $B + \lambda I$ is SPD, then determine the decomposition $R^T R = B + \lambda I$ and go to Step 4. Otherwise determine a vector $v \in R^n$, such that $\parallel v \parallel = 1$ and $v^T (B + \lambda I) v < 0$, set $\underline{\gamma} = \lambda - v^T (B + \lambda I) v$ and go to Step 2.

Step 4 Determine a vector $s \in R^n$ as the solution of the equation $R^T Rs + g = 0$. If $\| s \| > \overline{\delta}\Delta$, then set $\underline{\lambda} = \lambda$ and go to Step 6. If $\underline{\delta}\Delta \leq \| s \| \leq \overline{\delta}\Delta$, then terminate the computation. If $\| s \| < \underline{\delta}\Delta$ and $\lambda = 0$, then terminate the computation. If $\| s \| < \underline{\delta}\Delta$ and $\lambda \neq 0$, then set $\overline{\lambda} = \lambda$ and go to Step 5.

Step 5 Determine a vector $v \in R^n$ in such a way that it is a good approximation of the eigenvector of the matrix B corresponding to the eigenvalue $\underline{\lambda}(B)$ and, simultaneously, $\| v \| = 1$ and $v^T s \geq 0$ hold. Determine a number $\alpha \geq 0$ to $\| s + \alpha v \| = \Delta$ hold. If $\alpha^2 \| Rv \|^2 \leq (1 - \underline{\delta}^2)(\| Rs \|^2 + \lambda\Delta^2)$, then set $s := s + \alpha v$ and terminate the computation. Otherwise set $\underline{\gamma} = \lambda - \| Rv \|^2$ and go to Step 6.

Step 6 Determine a vector $v \in R^n$ as a solution of the equation $R^T v = s$ and set

$$\lambda := \lambda + \frac{\| s \|^2}{\| v \|^2} \left(\frac{\| s \| - \Delta}{\Delta} \right).$$

If $\lambda < \underline{\lambda}$, then set $\lambda = \underline{\lambda}$. If $\lambda > \overline{\lambda}$, then set $\lambda = \overline{\lambda}$. Go to Step 2.

3.4 Inexact trust region methods

For the determination of the trust region step, we can effectively use the conjugate gradient method applied on the minimization of the quadratic function

$$Q(s) = g^T s + \frac{1}{2} s^T B s$$

(We omit the index i). The conjugate gradient method uses the recurrence relations

$$s_1 = 0, \quad g_1 = g \quad p_1 = -g$$

and

$$\begin{gathered} q_i = Bp_i, \quad \alpha_i = \| g_i \|^2 / p_i^T q_i \\ s_{i+1} = s_i + \alpha_i p_i, \\ g_{i+1} = g_i + \alpha_i q_i, \quad \beta_i = \| g_{i+1} \|^2 / \| g_i \|^2, \\ p_{i+1} = -g_{i+1} + \beta_i p_i, \end{gathered} \tag{CG}$$

for $1 \leq i \leq n$.

Remark 40 It holds $g_i = Bs_i + g$.

Remark 41 The value $\alpha_i = \| g_i \|^2 / p_i^T q_i$ realizes the perfect line search, since

$$p_i^T g_{i+1} = p_i^T g_i + \alpha_i p_i^T q_i = - \| g_i \|^2 + (\| g_i \|^2 / p_i^T g_i) p_i^T q_i = 0.$$

We use the inductive step

$$p_i^T g_i = - \| g_i \|^2 + \beta_{i-1} p_{i-1}^T g_i = - \| g_i \|^2 .$$

Theorem 25 If we apply the conjugate gradient method on the quadratic function $Q(s)$ and if $p_i^T B p_i > 0$ holds for $1 \leq i \leq m$, then

$$\begin{aligned} Q(s_{i+1}) &\leq -\frac{1}{2} \| g \|^2 / \| B \|, \\ Q(s_{i+1}) &< Q(s_i), \\ \| s_{i+1} \| &> \| s_i \| \end{aligned}$$

for $1 \leq i \leq m$.

Proof From the proof of Theorem 10 it follows that

$$p_j^T B p_i = 0 \qquad \forall 1 \le j < i \le m,$$

$$g_j^T g_i = 0 \qquad \forall 1 \le j < i \le m+1,$$

$$p_j^T g_i = 0 \qquad \forall 1 \le j < i \le m+1.$$

If we use (CG), we obtain

$$\begin{aligned} Q(s_{i+1}) &= g^T(s_i + \alpha_i p_i) + \frac{1}{2}(s_i + \alpha_i p_i)^T B (s_i + \alpha_i p_i) = \\ &= Q(s_i) + \alpha_i g^T p_i + \alpha_i s_i^T B p_i + \frac{1}{2}\alpha_i^2 p_i^T B p_i = \\ &= Q(s_i) + \alpha_i g_i^T p_i + \frac{1}{2}\alpha_i^2 p_i^T B p_i = \\ &= Q(s_i) - \frac{\| g_i \|^4}{p_i^T B p_i} + \frac{1}{2}\frac{\| g_i \|^4}{p_i^T B p_i} = \\ &= Q(s_i) - \frac{1}{2}\frac{\| g_i \|^4}{p_i^T B p_i} < Q(s_i). \end{aligned}$$

Furthermore, it holds

$$\begin{aligned} \| s_{i+1} \|^2 &= (s_i + \alpha_i p_i)^T (s_i + \alpha_i p_i) = \| s_i \|^2 + \alpha_i^2 \| p_i \|^2 + 2\alpha_i s_i^T p_i = \\ &= \| s_i \|^2 + \alpha_i^2 \| p_i \|^2 + 2\alpha_i \sum_{j=1}^{i-1} \alpha_j p_j^T p_i = \\ &= \| s_i \|^2 + \alpha_i^2 \| p_i \|^2 + 2\alpha_i \sum_{j=1}^{i-1} \alpha_j \frac{\| p_j \|^2}{\| g_j \|^2} \| g_i \|^2 > \| s_i \|^2 . \end{aligned}$$

since

$$\begin{aligned} p_j^T p_i &= p_j^T(-g_i + \beta_{i-1} p_{i-1}) = \beta_{i-1} p_j^T p_{i-1} = \\ &= \left(\prod_{k=j}^{i-1} \beta_k \right) \| p_j \|^2 = (\| g_i \|^2 / \| g_j \|^2) \| p_j \|^2 \end{aligned}$$

Since

$$s_2 = s_1 + \frac{\| g_1 \|^2}{p_1^T B p_1} p_1 = -\frac{\| g \|^2}{g^T B g} g,$$

the part (b) of the proof of Theorem 23 implies

$$-Q(s_2) \ge \frac{1}{2} \| g \|^2 / \| B \|,$$

which together with $Q(s_{i+1}) < Q(s_i)$ for $1 \le i \le m$ gives $Q(s_{i+1}) < -(1/2) \| g \|^2 / \| B \|$.

Remark 42 If $p_i^T B p_i \le 0$, then

$$Q(s_i + \alpha_i p_i) = Q(s_i) + \alpha_i g_i^T p_i + \frac{1}{2}\alpha_i^2 p_i^T B p_i \le Q(s_i) - \alpha_i \| g_i \|^2 < Q(s_i)$$

for an arbitrary value $\alpha_i \ge 0$. If $\| s_i \| \le \Delta$ a $p_i^T B p_i \le 0$, then we determine the number $\alpha_i \ge 0$ so that $\| s_i + \alpha_i p_i \| = \Delta$ holds and set $s = s_i + \alpha_i p_i$. Using Theorem 25, we obtain $Q(s) \le -(1/2) \| g \|^2 / \| B \|$ for $i \ge 2$. From the part (c) of the proof of Theorem 23 it holds also for $i = 1$.

Remark 43 The number $\alpha_i \ge 0$, for which $\| s_i + \alpha_i p_i \| = \Delta$ holds, is determined by the formula

$$\alpha_i = -p_i^T s_i + \sqrt{(p_i^T s_i)^2 + \Delta^2 - \| s_i \|^2}.$$

Remark 44 The conjugate gradient method generates a path in the trust region, i.e. the a curve $s(t) \in R^n$, such that

$$\frac{d \parallel s(t) \parallel}{dt} > 0,$$
$$\frac{dQ(s(t))}{dt} < 0.$$

Theorem 25 is a basis for the following simple Algorithm:

Algorithm 5 Data $0 < \omega < 1$, $0 < \Delta$, $m \geq n$.

Step 1 Set $s = 0$, $r = -g$, $\sigma = \parallel r \parallel^2$, $p = r$ and $k = 1$.

Step 2 Set $\rho = \sigma$, compute the vector $q = Bp$ and the number $\tau = p^T q$. If $\tau \leq 0$, then determine a number $\alpha \geq 0$ such that $\parallel s + \alpha p \parallel = \Delta$, set $s := s + \alpha p$ and terminate the computation.

Step 3 Set $\alpha = \rho/\tau$. If $\parallel s + \alpha p \parallel \geq \Delta$, then determine a number $\alpha \geq 0$ such that $\parallel s + \alpha p \parallel = \Delta$, set $s := s + \alpha p$ and terminate the computation.

Step 4 Set $s := s + \alpha p$, $r := r - \alpha q$ and $\sigma = \parallel r \parallel^2$. If $\sigma \leq \omega^2 \parallel g \parallel^2$ or $k \geq m$, then terminate the computation.

Step 5 Set $\beta = \sigma/\rho$, $p := r + \beta p$, $k := k + 1$ and go to Step 2.

We usually choose $m = n + 3$.

3.5 Dog-leg methods

A disadvantage of trust region methods with the optimal locally constrained step is the necessity of solving the n-dimensional problem ($\overline{\text{T}}1$) which requires repeated solution of the linear system $(B_i + \lambda I)s_i(\lambda) + g_i = 0$. This system is solved (on average) 2-3 times in every iteration. For this reason, the n-dimensional problem ($\overline{\text{T}}1$) is frequently replaced by the simpler problem

$$s_i = \arg \min_{\|s(\alpha,\beta)\| \leq \Delta_i} Q_i(s(\alpha, \beta)) \qquad (\tilde{\text{T}}1),$$

where

$$s(\alpha, \beta) = \alpha g_i + \beta B_i^{-1} g_i.$$

The problem ($\tilde{\text{T}}1$) has dimension 2 and the linear system with the matrix B_i is solved only once (for determination of the vector $B_i^{-1} g_i$). The vector s_i, obtained by solving the problem ($\tilde{\text{T}}1$), satisfies conditions (T1) with $\overline{\omega} = 0$ and $\underline{\sigma} = 1/2$. Using this way, we obtain methods which converge almost as good as methods with the optimal locally constrained step. Computational experience shows that the efficiency of methods based on projections into the subspace spanned by the vectors g_i and $B_i^{-1} g_i$ is not changed too much if we replace the exact solution of the problem ($\tilde{\text{T}}1$) by a special approximate choice of the coefficients α and β, which gives the called dog leg method. In this section, we will consider a more general algorithm, depending on the integer m, which reduces to the dog-leg method when $m = 1$.

Theorem 26 Let the assumptions of Theorem 25 be satisfied for $m \geq 1$, where $\parallel s_{m+1} \parallel < \Delta$ and $Bs_{m+1} + g \neq 0$. Let $s_n \in R^n$ be a vector such that $Bs_n + g = 0$. Then

$$\frac{dQ(s_{m+1} + \alpha(s_n - s_{m+1}))}{d\alpha} = (1 - \alpha)(s_n - s_{m+1})^T g_{m+1}.$$

Proof Since

$$\begin{aligned} Q(s_{m+1} + \alpha(s_n - s_{m+1})) &= g^T(s_{m+1} + \alpha(s_n - s_{m+1})) + \\ &+ \frac{1}{2}(s_{m+1} + \alpha(s_n - s_{m+1}))^T B(s_{m+1} + \alpha(s_n - s_{m+1})), \end{aligned}$$

we obtain

$$
\begin{aligned}
\frac{dQ(s_{m+1}+\alpha(s_n-s_{m+1}))}{d\alpha} &= (s_n-s_{m+1})^T g+(s_n-s_{m+1})^T B(s_{m+1}+\alpha(s_n-s_{m+1})) = \\
&= (s_n-s_{m+1})^T B(s_{m+1}-s_n+\alpha(s_n-s_{m+1})) = \\
&= (1-\alpha)(s_n-s_{m+1})^T B(s_{m+1}-s_n) = \\
&= (1-\alpha)(s_n-s_{m+1})^T (Bs_{m+1}+g) = \\
&= (1-\alpha)(s_n-s_{m+1})^T g_{m+1}.
\end{aligned}
$$

Theorem 26 is a basis for the following algorithm, which is a combination of Algorithm 5 and the dog-leg method.

Algorithm 6 Data $0<\Delta$, $m<n$.

Step 1 As in Algorithm 5.

Step 2 As in Algorithm 5.

Step 3 As in Algorithm 5.

Step 4 Set $s:=s+\alpha p$, $r:=r-\alpha q$ and $\sigma=\| r \|^2$. If $k<m$, then set $\beta=\sigma/\rho$, $p:=r+\beta p$, $k:=k+1$ and go to Step 2.

Step 5 Solve the linear system $Bs^*+g=0$. If $(s^*-s)^T r>0$, then determine a number $\alpha\geq 0$ such that $\| s+\alpha(s^*-s) \|=\Delta$. If $\alpha\geq 1$, then set $s:=s^*$ and terminate the computation. If $\alpha<1$, then set $s:=s+\alpha(s^*-s)$ and terminate the computation. If $(s^*-s)^T r\leq 0$, then determine a number $\alpha\geq 0$ such that $\| s-\alpha(s^*-s) \|=\Delta$, set $s:=s-\alpha(s^*-s)$ and terminate the computation.

We usually choose $m\leq 3$. For $m=1$ we obtain the dog-leg method.

3.6 Decomposition of symmetric indefinite matrices

1) The Gill-Murray decomposition of the matrix B has the form

$$R^T R = B+E,$$

where R is a regular upper triangular matrix and E is a positive semidefinite diagonal matrix ($E=0$ if B is sufficiently positive definite). At the start of i-th elimination step we have the matrix

$$
\begin{bmatrix}
R_{(i-1),(i-1)}, & R_{(i-1),i}, & R_{(i-1),(n-i)} \\
*, & B_{ii}^{(i-1)}, & B_{i,(n-i)}^{(i-1)} \\
*, & *, & B_{(n-i),(n-i)}^{(i-1)}
\end{bmatrix},
$$

where the upper index in parantheses denotes the number of applied elimination steps and lower indices in parantheses denote submatrices with $(i-1)$ or $(n-i)$ rows or columns, respectively. The elimination step has the following form

$$\gamma_i = \max_{i<j\leq n}(|B_{i,j}^{(i-1)}|),$$

$$\rho_i^2 = \max\left(|B_{ii}^{(i-1)}|, \frac{\gamma_i^2}{\beta^2}, \delta^2\right),$$

$$R_{ii}=\rho_i,$$

$$R_{i,(n-i)} = B_{i,(n-i)}^{(i-1)}/R_{ii},$$

$$B_{(n-i),(n-i)}^{(i)} = B_{(n-i),(n-i)}^{(i-1)} - R_{i,(n-i)}^T R_{i,(n-i)},$$

where δ is a small number and $\beta > \sqrt{\| B \|}$). This process differs from the Choleski decomposition only by the possibility that $\rho_i^2 \neq B_{ii}^{(i-1)}$ can hold. Using a more detailed recognition of the given relations, we can prove that the diagonal elements of the matrix E are expressed in the form

$$E_{ii} = \rho_i^2 - B_{ii}^{(i-1)} = \rho_i^2 + R_{i,(n-i)} R_{i,(n-i)}^T - B_{ii}$$

where B_{ii}, $1 \le i \le n$, are the diagonal elements of the original matrix B.

Theorem 27 Let $R^T R = B + E$ be the Gill-Murray decomposition with $\delta = 0$ and $\beta > \sqrt{\| B \|}$. Let

$$B_{kk}^{(k-1)} = \min_{1 \le i \le n} B_{ii}^{(i-1)}$$

and let $v \in R^n$ be a vector determined as the solution of the equation $Rv = e_k$ (e_k is k-th column of the unit matrix). If the matrix B is not positive semidefinite, then

$$v^T B v = \frac{B_{kk}^{(k-1)}}{\rho_k^2} < 0.$$

Proof Equation $Rv = e_k$ implies that $v_k = 1/\rho_k$. Therefore, it holds

$$\begin{aligned} v^T B v &= v^T (B + E) v - v^T E v \le v^T R^T R v - v_k^2 E_{kk} = \\ &= e_k^T e_k - E_{kk}/\rho_k^2 = \frac{\rho_k^2 - E_{kk}}{\rho_k^2} = \frac{B_{kk}^{(k-1)}}{\rho_k^2}. \end{aligned}$$

If the matrix B is not positive semidefinite, then an index $1 \le i \le n$ has to exist, such that $E_{ii} \neq 0$ or, equivalently, $\rho_i^2 \neq B_{ii}^{(i-1)}$. Two cases can arise. Either $\rho_i^2 = |B_{ii}^{(i-1)}| \neq B_{ii}^{(i-1)}$, so that $B_{ii}^{(i-1)} < 0$ and hence $B_{kk}^{(k-1)} < 0$, or $\rho_i^2 = \gamma_i^2/\beta^2$. In the second case, an index $i < j \le n$ has to exist, such that $\gamma_i = |B_{ij}^{(i-1)}|$, so that

$$|R_{ij}| = \frac{|B_{ij}^{(i-1)}|}{\rho_i} = \frac{\gamma_i}{\gamma_i/\beta} = \beta,$$

which gives

$$B_{ii}^{(i-1)} = \rho_i^2 - E_{ii} = B_{ii} - R_{i,(n-i)} R_{i,(n-1)}^T \le B_{ii} - \beta^2 < \| B \| - \| B \| = 0.$$

2) The Bunch-Parlett decomposition of the matrix B has the form

$$LDL^T = PBP^T,$$

where

$$L = \begin{bmatrix} I, & 0, & \dots, & 0 \\ L_{21}, & I, & \dots, & 0 \\ \vdots & \vdots & \vdots & \vdots \\ L_{n1}, & L_{n2}, & \dots, & I \end{bmatrix}, \quad D = \begin{bmatrix} D_{11}, & 0, & \dots, & 0 \\ 0, & D_{22}, & \dots, & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0, & 0, & \dots, & D_{nn} \end{bmatrix}$$

Therefore, L is a lower triangular matrix with unit diagonal blocks and D is a block diagonal matrix (blocks have dimension either 1×1 or 2×2). At the start of i-th elimination step we have the matrix

$$\begin{bmatrix} D_{11}, & L_{12}, & \dots, & L_{1,i-1}, & L_{1,(m-i+1)} \\ *, & D_{22}, & \dots, & L_{2,i-1}, & L_{2,(m-i+1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ *, & *, & \dots, & D_{i-1,i-1}, & L_{i-1,(m-i+1)} \\ *, & *, & \dots, & *, & B^{(i-1)} \end{bmatrix}.$$

The elimination step has the following form

$$\beta_i = \max_k |B_{kk}^{(i-1)}|,$$

$$\gamma_i = \max_{k,l} |B_{kl}^{(i-1)}|,$$

$$\alpha_i = \beta_i / \gamma_i .$$

If $\alpha_i \geq \left(\sqrt{17}+1\right)/8$, then we choose 1×1 block, otherwise we choose 2×2 block. It is necessary to use permutations (the pivot block with indices k and l is transferred into the upper left corner of the matrix $B^{(i-1)}$). Then the transformation

$$B^{(i-1)} \rightarrow \begin{bmatrix} D_{ii}, & L_{i,(m-i)} \\ *, & B^{(i)} \end{bmatrix}$$

is realized, where

$$D_{ii} = B_{ii}^{(i-1)},$$

$$L_{i,(m-i)} = D_{ii}^{-1} B_{i,(m-i)}^{(i-1)},$$

$$B^{(i)} = B_{(m-i),(m-i)}^{(i-1)} - L_{i,(m-i)}^T B_{i,(m-i)}^{(i-1)} .$$

Theorem 28 Let $LDL^T = PBP^T$ be the Bunch-Parlett decomposition. Let $u_i = 0$, if $\underline{\lambda}(D_{ii}) \geq 0$, and let u_i be the normalised eigenvector corresponding to $\underline{\lambda}(D_{ii})$, if $\underline{\lambda}(D_{ii}) < 0$. Let $L^T P v = u$, where $u^T = [u_1, \ldots, u_m]$. If the matrix B is not positive semidefinite, then

$$v^T B v = \sum_{\underline{\lambda}(D_{ii})<0} \underline{\lambda}(D_{ii}) < 0.$$

Proof Equation $L^T P v = u$ implies

$$v^T B v = v^T P^T L D L^T P v = u^T D u = \sum_{i=1}^{m} u_i^T D_{ii} u_i = \sum_{\underline{\lambda}(D_{ii})<0} \underline{\lambda}(D_{ii}).$$

If the matrix B is not positive semidefinite, at least one block D_{kk} of the matrix D exists, which is not positive semidefinite, so that $\underline{\lambda}(D_{kk}) < 0$. Therefore, it holds

$$v^T B v = \sum_{\underline{\lambda}(D_{ii})<0} \underline{\lambda}(D_{ii}) \leq \underline{\lambda}(D_{kk}) < 0.$$

3.7 Newton method

The Newton method utilizes the matrices $B_i = G(x_i), i \in N$, so that (F3) implies $\| B_i \| = \| G(x_i) \| \leq \overline{G}, i \in N$.

Theorem 29 Let the assumptions (F1) and (F3) be satisfied. Then the trust region Newton method is globally convergent. If in addition the conditions (F4) and (F5) are satisfied and $x_i \rightarrow x^*$ and $\omega_i(s_i) \rightarrow 0$ hold, then rate of convergence is Q-superlinear.

Proof Global convergence follows immediately from Theorem 21 ($\| B_i \| \leq \overline{G}, i \in N$ holds). If $x_i \rightarrow x^*$, then $B_i = G(x_i) \rightarrow G(x^*)$, or

$$\frac{\| (G^* - B_i) s_i \|}{\| s_i \|} \leq \| G^* - B_i \| \rightarrow 0,$$

which together with $\omega_i(s_i) \rightarrow 0$ implies Q-superlinear convergence (Theorem 22).

The following realizations of the Newton method are most commonly used

1) Inexact Newton method ($\omega_i(s_i) > 0$). If (F3)-(F5) and $\omega_i(s_i) \to 0$ hold, then this realization is Q-superlinearly convergent (system $B_i s_i + g_i = 0$ is solved inexact by the conjugate gradient method $\Rightarrow$ less than $O(n^3)$ operations per iteration, which is advantageous for large-scale problems).

2) Newton method with the optimal locally constrained step. For this realization, a very strong statement holds:

Proposition 6 Let the conditions (F1)-(F3) be satisfied. Let x_i, $i \in N$ be a sequence determined by the Newton method with the optimal locally constrained step. Then a cluster point $x^* \in R^n$ of the sequence x_i, $i \in N$ exists, such that $g(x^*) = 0$ and $G(x^*) \geq 0$. If in addition the point $x^* \in R^n$ satisfies sufficient conditions for local minimum ($g(x^*) = 0$ and $G(x^*) > 0$), then $x^* \in R^n$ is an unique cluster point of the sequence x_i, $i \in N$, and the sequence x_i, $i \in N$, converges Q-superlinearly to the point $x^* \in R^n$.

The following table shows a comparison of several realizations of the Newton method (L - line search realization, T - trust region realization, G - Gill-Murray decomposition, B - Bunch-Parlett decomposition) by means of 15 test functions with 20 variables (the total numbers of iterations NIT, function evaluations NFV and gradient evaluations NFG are given together with the number of failures and the total computational time).

Method	NIT - NFV - NFG	fail	time
L - G	312 - 508 - 508	1	7.63
L - B	342 - 488 - 488	1	16.03
T - G (optimal)	281 - 321 - 296	-	4.61
T - B (dog-leg)	292 - 325 - 307	-	10.38

3.8 Gauss-Newton method for sum of squares

Suppose that the objective function $F(x)$ has the form

$$F(x) = \frac{1}{2} f^T(x) f(x) = \frac{1}{2} \sum_{k=1}^{m} f_k^2(x),$$

where $f_k : R^n \to R$, $1 \leq k \leq m$, are twice continuously differentiable functions. Then

$$g(x) = J^T(x) f(x) = \sum_{k=1}^{m} f_k(x) g_k(x),$$

$$G(x) = J^T(x) J(x) + C(x) = \sum_{k=1}^{m} g_k(x) g_k^T(x) + \sum_{k=1}^{m} f_k(x) G_k^T(x).$$

The Gauss-Newton method is obtained from the Newton method if the term $C(x_i)$ in the relation for $G(x_i)$ is omitted, so that

$$B_i = J_i^T J_i = \sum_{k=1}^{m} g_k(x_i) g_k^T(x_i)$$

Reasons:

1) Zero residual problems ($F(x^*) = 0$). From $x_i \to x^*$ it follows that $F(x_i) \to F(x^*) = 0$ and, therefore, $f_k(x_i) \to 0 \ \forall 1 \leq k \leq m$. If $\| G_k(x) \| \leq \overline{G}$, then also

$$\| C(x_i) \| = \| \sum_{k=1}^{m} f_k(x_i) G_k(x_i) \| \leq \overline{G} \sum_{k=1}^{m} |f_k(x_i)| \to 0$$

and, therefore, $\| G(x_i) - B_i \| = \| C(x_i) \| \to 0$, which implies Q-superlinear rate of convergence.

2) Linearization. It holds

$$\begin{aligned} F(x_i + s) &= \frac{1}{2} f^T(x_i + s) f(x_i + s) \approx \frac{1}{2}(f(x_i) + J(x_i)s)^T (f(x_i) + J(x_i)s) = \\ &= \frac{1}{2} f^T(x_i) f(x_i) + f^T(x_i) J(x_i) s + \frac{1}{2} s^T J^T(x_i) J(x_i) s \end{aligned}$$

so that

$$F(x_i + s) - F(x_i) \approx g^T(x_i)s + \frac{1}{2} s^T B_i s,$$

which is a local approximation with the matrix $B_i = J_i^T J_i$.

In the subsequent considerations, it is necessary to slightly modify conditions which the objective function $F : R^n \to R$ has to satisfy. The condition (F1) is always satisfied, since $F(x) \geq 0 \ \forall x \in R^n$. The condition (F3) will be replaced by the condition

$$\| G_k(x) \| \leq \overline{G} \qquad (\overline{\text{F3}})$$

$\forall x \in R^n$, $\forall 1 \leq k \leq m$. From (F2) and ($\overline{\text{F3}}$), the boundedness of gradients and function values

$$\| g_k(x) \| \leq \overline{g},$$

$$|f_k(x)| \leq \overline{f}$$

$\forall x \in \mathcal{L}(F(x_1))$, $\forall 1 \leq k \leq m$, follows and, therefore, (F3) holds.

Theorem 30 Let the conditions (F2) and ($\overline{\text{F3}}$) be satisfied. Then the trust region Gauss-Newton method is globally convergent. If, in addition, the conditions (F4) and (F5) are satisfied and $x_i \to x^*$, $F(x^*) = 0$ together with $\omega_i(s_i) \to 0$ holds, then the rate of convergence is Q-superlinear.

Proof From (F2) and ($\overline{\text{F3}}$) it follows that $\| G_k(x) \| \leq \overline{G}$, $\| g_k(x) \| \leq \overline{g}$, $|f_k(x)| \leq \overline{f} \ \forall x \in R^n$, $\forall 1 \leq k \leq m$. Therefore, it holds

$$\| G(x) \| \leq \sum_{k=1}^{m} \| g_k(x) \|^2 + \sum_{k=1}^{m} |f_k(x)| \, \| G_k(x) \| \leq m\overline{g}^2 + m\overline{f}\overline{G}$$

(condition (F3)) and

$$\| B_i \| = \| \sum_{k=1}^{m} g_k(x_i) g_k^T(x_i) \| \leq \sum_{k=1}^{m} \| g_k(x_i) \|^2 \leq m\overline{g}^2,$$

so that, by Theorem 21 the Gauss-Newton method is globally convergent. As it was already proved, $F(x_i) \to F(x^*) = 0$ implies $B_i \to G(x_i) \to G(x^*)$, or

$$\frac{\| (G^* - B_i)s_i \|}{\| s_i \|} \leq \| G^* - B_i \| \to 0,$$

which together with $\omega_i(s_i) \to 0$ implies Q-superlinear convergence (Theorem 22).

The direction vector corresponding to the Gauss-Newton method can be determined by three different ways:

1) Solving the normal system of equations. The equation $B_i s_i + g_i = 0$ has the form

$$J_i^T J_i s_i + J_i^T f_i = 0. \qquad (\text{NE})$$

2) Solving the linearized least squares problem (overdetermined system of linear equations). The linearized problem has the form tvar

$$J_i s_i + f_i \approx 0. \tag{OE}$$

The QR-decomposition $J_i = Q_i \begin{bmatrix} R_i \\ 0 \end{bmatrix}$, where $Q_i^T Q_i = I$, can be used, so that

$$Q_i^T J_i s_i = \begin{bmatrix} R_i \\ 0 \end{bmatrix} s_i = Q_i^T f_i,$$

where R_i is an upper triangular matrix. The QR-decomposition is stable and it is possible to determine the pseudorank of the matrix J_i and, therefore, decrease the dimension of the problem solved. For trust region realization, the system

$$(J_i^T J_i + \lambda I)s + J_i^T f_i = 0$$

can be replaced by the linear least squares problem

$$\begin{bmatrix} J_i \\ \sqrt{\lambda} I \end{bmatrix} s + \begin{bmatrix} f_i \\ 0 \end{bmatrix} \approx 0.$$

3) Solving the augmented system of equations. Denote $r_i = -(J_i s_i + f_i)$. The direction vector s_i is looked for to satisfy the equation $J_i^T r_i = 0$. This together gives

$$\begin{bmatrix} I & J_i \\ J_i^T & 0 \end{bmatrix} \begin{bmatrix} r_i \\ s_i \end{bmatrix} + \begin{bmatrix} f_i \\ 0 \end{bmatrix} = 0, \tag{SE}$$

which is a system of $m+n$ equations with a symmetric indefinite matrix. This way is suitable for sparse or weighted problems. If

$$F(x) = \frac{1}{2} f^T(x) W f(x),$$

where W is a weighting matrix, then the normal system of equations has the form

$$J_i^T W J_i s_i + J_i^T W f_i = 0$$

and denoting $r_i = -W(J_i s_i + f_i)$, we get

$$\begin{bmatrix} W^{-1} & J_i \\ J_i^T & 0 \end{bmatrix} \begin{bmatrix} r_i \\ s_i \end{bmatrix} + \begin{bmatrix} f_i \\ 0 \end{bmatrix} = 0,$$

so that some of weight coefficients can be infinite (constrained least squares problems).

3.9 Hybrid methods for sum of squares

The Gauss-Newton method is very efficient for zero residual problems, however, it can fail in the case of large residual problems. Therefore, the following strategy seems to be advantageous:

$F_i \to F^* = 0 \Rightarrow$ The Gauss-Newton method,
$F_i \to F^* > 0 \Rightarrow$ The BFGS variable metric method.

Lemma 8 Let $F_i \to F^* = 0$ Q-superlinearly. Then

$$\lim_{i \to \infty} \frac{F_i - F_{i+1}}{F_i} = 1.$$

Let $F_i \to F^* > 0$. Then

$$\lim_{i \to \infty} \frac{F_i - F_{i+1}}{F_i} = 0.$$

Proof If $F_i \to F^* = 0$ Q-superlinearly, then

$$\lim_{i\to\infty} \frac{F_i - F_{i+1}}{F_i} = 1 - \lim_{i\to\infty} \frac{F_{i+1} - F^*}{F_i - F^*} = 1 - 0 = 1.$$

If $F_i \to F^* > 0$, then

$$\lim_{i\to\infty} \frac{F_i - F_{i+1}}{F_i} = \frac{1}{F^*} \lim_{i\to\infty} (F_i - F_{i+1}) = 0.$$

The most simple hybrid method has the following form. Let $B_1 = J_1^T J_1$. If $i > 1$ and $(F_i - F_{i+1})/F_i > \underline{\vartheta}$, then set

$$B_{i+1} = J_{i+1}^T J_{i+1}.$$

If $i > 1$ and $(F_i - F_{i+1})/F_i \le \underline{\vartheta}$, then set

$$B_{i+1} = B_i + \frac{y_i y_i^T}{y_i^T d_i} - \frac{B_i d_i (B_i d_i)^T}{d_i^T B_i d_i},$$

where $y_i = g_{i+1} - g_i$ and $d_i = x_{i+1} - x_i$. Usually $\underline{\vartheta} = 0.01$ for line search methods and $\underline{\vartheta} = 0.0001$ for trust region methods.

Some other hybrid methods are described in Section 4.7. The following table shows a comparison of several methods for minimization of sum of squares (L - line search method, T - trust region method, GN - Gauss-Newton method, VM - variable metric method, GN+VM - hybrid method) by means of 30 test problems with 2-12 variables (the total numbers of iterations NIT, function evaluations NFV and gradient eveluations NFG are given together with the number of failures and the total computational time).

Method	IT - IF - IG	fail	time
S - GN	1917 - 2974 - 2974	3	9.67
S - VM	1543 - 3256 - 3256	2	5.71
S - GN+VM	635 - 1037 - 1037	-	4.12
T - GN	605 - 748 - 634	-	2.64
T - VM	2155 - 2542 - 2183	2	7.58
T - GN+VM	536 - 664 - 565	-	2.58

4 Methods for large-scale sparse or partitioned problems

Large-scale problems cannot be solved by methods which require dense matrices. Some special methods have to be used in the large-scale case:

- Variable metric methods with limited storage.
- Difference version of the inexact Newton method.
- Special methods for sparse problems (sparse Newton method, sparse variable metric methods).
- Special methods for partitioned problems (partitioned Newton method, partitioned variable metric methods).
- Sparse modifications of the Gauss-Newton method for sum of squares.

4.1 Variable metric methods with limited storage

Variable metric methods with limited storage are based on the use of a limited number of steps of the BFGS method.

Lemma 9 The BFGS update can be written in the form

$$H_+ = \gamma V^T H V + \frac{\rho}{b} dd^T,$$

where

$$V = I - \frac{1}{b} y d^T$$

and where $y = g_+ - g$, $d = x_+ - x$ and $a = y^T H y$, $b = y^T d$.

Proof Using straighforvard manipulations, we get

$$H_+ = \gamma \left(H - \frac{1}{b}(H y d^T + d(Hy)^T) + \frac{a}{b}\frac{1}{b} dd^T + \frac{\rho}{\gamma}\frac{1}{b} dd^T \right).$$

The same result can be obtained by a simple modification of the relation

$$H_+ = \gamma \left(H + \frac{\rho}{\gamma}\frac{1}{b} dd^T - \frac{1}{a} Hy(Hy)^T + \frac{1}{a}\left(\frac{a}{b}d - Hy\right)\left(\frac{a}{b}d - Hy\right)^T \right)$$

corresponding to the BFGS method.

Definition 24 Let $\overline{m} > 0$ and $m = \min(\overline{m}, i-1)$. We say that a basic optimization method is the $\overline{m}$-step limited storage BFGS method if

$$s_i = -H_i^i g_i,$$

where $H_{i-m}^i = I$ and

$$H_{j+1}^i = \gamma_j^i V_j^T H_j^i V_j + \frac{\rho_j}{b_j} d_j d_j^T$$

for $i-m \le j \le i-1$. Here $\gamma_{i-m}^i = b_{i-1}/a_{i-1}$ and $\gamma_j^i = 1$ for $i-m < j \le i-1$.

Theorem 31 For the $\overline{m}$-step limited storage BFGS method it holds

$$H_{j+1}^i = \frac{b_{i-1}}{a_{i-1}} \left(\prod_{k=i-m}^{j} V_k \right)^T \left(\prod_{k=i-m}^{j} V_k \right) + \sum_{l=i-m}^{j} \frac{\rho_l}{b_l} \left(\prod_{k=l+1}^{j} V_k \right)^T d_l d_l^T \left(\prod_{k=l+1}^{j} V_k \right).$$

Proof (By induction) For $j = i-m$ it holds (it suffices to substitute $\gamma_{i-m}^i = b_{i-1}/a_{i-1}$ and $H_{i-m}^i = I$ into the BFGS update). The inductive step:

$$\begin{aligned} H_{j+1}^i &= V_j^T H_j^i V_j + \frac{\rho_j}{b_j} d_j d_j^T = \frac{b_{i-1}}{a_{i-1}} V_j^T \left(\prod_{k=i-m}^{j-1} V_k \right)^T \left(\prod_{k=i-m}^{j-1} V_k \right) V_j + \\ &+ \sum_{l=i-m}^{j-1} \frac{\rho_l}{b_l} V_j^T \left(\prod_{k=l+1}^{j-1} V_k \right)^T d_l d_l^T \left(\prod_{k=l+1}^{j-1} V_k \right) V_j + \frac{\rho_j}{b_j} d_j d_j^T = \\ &= \frac{b_{i-1}}{a_{i-1}} \left(\prod_{k=i-m}^{j} V_k \right)^T \left(\prod_{k=i-m}^{j} V_k \right) + \sum_{l=i-m}^{j} \frac{\rho_l}{b_l} \left(\prod_{k=l+1}^{j} V_k \right)^T d_l d_l^T \left(\prod_{k=l+1}^{j} V_k \right). \end{aligned}$$

Theorem 31 shows that the matrix H_i^i can be obtained by means of $2m$ vectors d_j, y_j, $i-m \le j \le i-1$, without storing matrices H_j^i, $i-m \le j \le i-1$. However, the matrix H_i^i may not be constructed explicitly, it suffices to compute the vector $s_i = -H_i^i g_i$, which is realized by two recurrence relations (the Strang formula). First, the vectors

$$u_j = -\left(\prod_{k=j}^{i-1} V_k\right) g_i,$$

where $i-m \le j \le i-1$, are computed using a backward recurrence. Since

$$u_j = V_j u_{j+1} = \left(I - \frac{1}{b_j} y_j d_j^T\right) u_{j+1} = u_{j+1} - \frac{d_j^T u_{j+1}}{b_j} y_j$$

for $i-m \le j \le i-1$, where $u_i = -g_i$, we can write

$$u_i = -g_i$$

and

$$\sigma_j = d_j^T u_{j+1}/b_j,$$

$$u_j = u_{j+1} - \sigma_j y_j \tag{R1}$$

for $i-m \le j \le i-1$. Secondly, the vectors

$$v_{j+1} = \frac{b_{i-1}}{a_{i-1}}\left(\prod_{k=i-m}^{j} V_k\right)^T u_{i-m} + \sum_{l=i-m}^{j} \frac{\rho_l}{b_l}\left(\prod_{k=l+1}^{j} V_k\right)^T d_l d_l^T u_{l+1},$$

where $i-m \le j \le i-1$, are computed using a forward recurrence. Since

$$v_{j+1} = V_j^T v_j + \frac{\rho_j}{b_j} d_j d_j^T u_{j+1} = \left(I - \frac{1}{b_j} d_j y_j^T\right) v_j + \rho_j \sigma_j d_j = v_j + (\rho_j \sigma_j - y_j^T v_j / b_j) d_j,$$

where $v_{i-m} = (b_{i-1}/a_{i-1}) u_{i-m}$, we can write

$$v_{i-m} = (b_{i-1}/a_{i-1}) u_{i-m}$$

and

$$v_{j+1} = v_j + (\rho_j \sigma_j - y_j^T v_j) d_j \tag{R2}$$

for $i-m \le j \le i-1$. Finally we set $s_i = v_i$.

Remark 45 It is only necessary to store the numbers σ_j, $i-m \le j \le i-1$. All vectors u_j, v_j, $i-m \le j \le i$ are stored on the same place as the vector s_i. In summary, we require to store $2m+3$ vectors (d_j, y_j, $i-m \le j \le i-1$ and 3 vectors for the basic optimization method) and we use $O(mn)$ multiplications.

Proposition 7 (Quadratic termination). Let x_i, $i \in N$ be a sequence generated by the m-step limited memory BFGS method with perfect line searches ($s_i^T g_{i+1} = 0\ \forall i \in N$ holds) applied to the strict convex quadratic function (Q). Then an index $k \le n$ exists, such that $g_{k+1} = 0$ and $x_{k+1} = x^*$.

The Strang formula (R1) and (R2) is oldest most simple realization of the m-step limited memory BFGS method. For some applications, the compact form derived below is more advantageous.

Lemma 10 Let $N = -M^{-1}$, where M is the matrix used in Theorem 15 (with $\rho = 1$ and $\gamma = 1$). Then

$$N = \begin{bmatrix} \frac{(\eta-1)b^2}{\eta a + (1-\eta)b}, & \frac{\eta ab}{\eta a + (1-\eta)b} \\ \frac{\eta ab}{\eta a + (1-\eta)b}, & a + \frac{\eta ab}{\eta a + (1-\eta)b} \end{bmatrix}. \tag{N}$$

Proof From expression of the matrix M (Theorem 15) it follows that

$$N = -M^{-1} = -\frac{1}{\det M} \begin{bmatrix} \frac{\eta-1}{a}, & \frac{\eta}{b} \\ \frac{\eta}{b}, & \frac{1}{b}(\eta\frac{a}{b}+1) \end{bmatrix}.$$

If we replace for $-\det M$ by the expression of m defined in Theorem 18, we obtain (N) after straightforward manipulations.

Remark 46 The DFP method uses $\eta = 0$, so that

$$N = \begin{bmatrix} -d^T y, & 0 \\ 0, & y^T H y \end{bmatrix}. \tag{ND}$$

The BFGS method uses $\eta = 1$, so that

$$N = \begin{bmatrix} 0, & d^T y \\ d^T y, & d^T y + y^T H y \end{bmatrix}. \tag{NB}$$

Lemma 11 Let B be a nonsingular square matrix and $\beta - b^T B^{-1} b \neq 0$. Then

$$[A, a] \begin{bmatrix} B, & b \\ b^T, & \beta \end{bmatrix}^{-1} \begin{bmatrix} A^T \\ a^T \end{bmatrix} = AB^{-1}A^T + (AB^{-1}b - a)(\beta - b^T B^{-1} b)^{-1}(b^T B^{-1} A^T - a^T).$$

Proof Realizing a matrix multiplication, we easily derive that

$$\begin{bmatrix} B, & b \\ b^T, & \beta \end{bmatrix}^{-1} = \begin{bmatrix} B^{-1} + B^{-1}b(\beta - b^T B^{-1} b)^{-1} b^T B^{-1}, & -B^{-1}b(\beta - b^T B^{-1} b)^{-1} \\ -(\beta - b^T B^{-1} b)^{-1} b^T B^{-1}, & (\beta - b^T B^{-1} b)^{-1} \end{bmatrix}.$$

Substituting this expression into the left hand side of the statement which we are proving and using a matrix multiplication, we obtain the required result.

In the sequel, we suppose that H_1 is a symmetric positive definite matrix and

$$H_{k+1} = H_k - [d_k, H_k y_k] N_k^{-1} \begin{bmatrix} d_k^T \\ y_k^T H_k \end{bmatrix} \tag{H}$$

holds for an arbitrary index $1 \leq k \leq m$, where N_k is a matrix specifying a particular variable metric method. We want to find the expression

$$H_{k+1} = H_1 - [D_k, H_1 Y_k] \overline{N}_k^{-1} \begin{bmatrix} D_k^T \\ Y_k^T H_1 \end{bmatrix}, \tag{$\overline{\mathrm{H}}$}$$

where $D_k = [d_1, \dots, d_k]$, $Y_k = [y_1, \dots, y_k]$ and where $\overline{N}_k$ is a symmetric matrix of order $2k$. We use the notation R_k for an upper triangular matrix of order k such that $(R_k)_{ij} = d_i^T y_j$, $i \leq j$, and $(R_k)_{ij} = 0$, $i > j$, and the notation C_k for a diagonal matrix of order k such that $(C_k)_{ij} = d_i^T y_j$, $i = j$, and $(C_k)_{ij} = 0$, $i \neq j$. To simplify the notation in the subsequent proofs, we omit the index k and replace the $k+1$ by $+$. For this convenience, we denote $D = [d_1, \dots, d_{k-1}]$, $Y = [y_1, \dots, y_{k-1}]$ and $R = R_{k-1}$, $C = C_{k-1}$, so that $D_k = [D, d]$, $Y_k = [Y, y]$ and

$$R_k = \begin{bmatrix} R, & D^T y \\ 0, & d^T y \end{bmatrix}, \quad R_k - C_k = \begin{bmatrix} R - C, & D^T y \\ 0, & 0 \end{bmatrix}.$$

Notice that using (H) and ($\overline{\mathrm{H}}$), we can write the induction step, used in the subsequent proofs, in the form

$$
\begin{aligned}
H_+ &= H - \left[d, H_1 y - [D, H_1 Y] \, \overline{N}^{-1} \begin{bmatrix} D^T \\ Y^T H_1 \end{bmatrix} y \right] \cdot \\
&\quad N^{-1} \begin{bmatrix} d^T \\ y^T H_1 - y^T [D, H_1 Y] \, \overline{N}^{-1} \begin{bmatrix} D^T \\ Y^T H_1 \end{bmatrix} \end{bmatrix} = \\
&= H - \left([d, H_1 y] - [D, H_1 Y] \, \overline{N}^{-1} \begin{bmatrix} 0, & D^T y \\ 0, & Y^T H_1 y \end{bmatrix} \right) \cdot \\
&\quad N^{-1} \left(\begin{bmatrix} d^T \\ y^T H_1 \end{bmatrix} - \begin{bmatrix} 0, & 0 \\ y^T D, & y^T H_1 Y \end{bmatrix} \overline{N}^{-1} \begin{bmatrix} D^T \\ Y^T H_1 \end{bmatrix} \right),
\end{aligned}
\tag{$*$}
$$

where $\overline{N} = \overline{N}_{k-1}$ and $N = N_k$.

Theorem 32 Let H_1 be a SPD matrix and let for an arbitrary index $1 \le k \le m$ the formula (H) hold, where the matrix N_k is defined by the relation (ND) (the DFP method). Then we can write

$$
H_{k+1} = H_1 - [D_k, H_1 Y_k] \begin{bmatrix} -C_k, & R_k - C_k \\ (R_k - C_k)^T, & Y_k^T H_1 Y_k \end{bmatrix}^{-1} \begin{bmatrix} D_k^T \\ Y_k^T H_1 \end{bmatrix}. \tag{$\overline{\mathrm{HD}}$}
$$

Proof For $k = 1$, ($\overline{\mathrm{HD}}$) is equivalent to (HD) (i.e. to (H) where the matrix N is determined by (ND)). The induction step: Suppose that ($\overline{\mathrm{HD}}$) holds for all indices less than k. At the stage k, the formula ($\overline{\mathrm{HD}}$) can be written in the form

$$
H_+ = H_1 - [D, H_1 Y, d, H_1 y] \begin{bmatrix} -C, & R - C, & 0, & D^T y \\ (R - C)^T, & Y^T H_1 Y, & 0, & Y^T H_1 y \\ 0, & 0, & -d^T y, & 0 \\ y^T D, & y^T H_1 Y, & 0, & y^T H_1 y \end{bmatrix}^{-1} \begin{bmatrix} D^T \\ Y^T H_1 \\ d^T \\ y^T H_1 \end{bmatrix}
$$

(after permutations which simplify the notation). Using Lemma 11 and denoting

$$
\overline{N} = \begin{bmatrix} -C, & R - C \\ (R - C)^T, & Y^T H_1 Y \end{bmatrix},
$$

we obtain

$$
\begin{aligned}
H_+ &= H_1 - [D, H_1 Y] \, \overline{N}^{-1} \begin{bmatrix} D^T \\ Y^T H_1 \end{bmatrix} - \\
&\quad \left([d, H_1 y] - [D, H_1 Y] \, \overline{N}^{-1} \begin{bmatrix} 0, & D^T y \\ 0, & Y^T H_1 y \end{bmatrix} \right) \cdot \\
&\quad \left(\begin{bmatrix} -d^T y, & 0 \\ 0, & y^T H_1 y \end{bmatrix} - \begin{bmatrix} 0, & 0 \\ y^T D, & y^T H_1 Y \end{bmatrix} \overline{N}^{-1} \begin{bmatrix} 0, & D^T y \\ 0, & Y^T H_1 y \end{bmatrix} \right)^{-1} \cdot \\
&\quad \left(\begin{bmatrix} d^T \\ y^T H_1 \end{bmatrix} - \begin{bmatrix} 0, & 0 \\ y^T D, & y^T H_1 Y \end{bmatrix} \overline{N}^{-1} \begin{bmatrix} D^T \\ Y^T H_1 \end{bmatrix} \right) = \\
&= H - \left([d, H_1 y] - [D, H_1 Y] \, \overline{N}^{-1} \begin{bmatrix} 0, & D^T y \\ 0, & Y^T H_1 y \end{bmatrix} \right) \cdot \\
&\quad \begin{bmatrix} -d^T y, & 0 \\ 0, & y^T H y \end{bmatrix}^{-1} \left(\begin{bmatrix} d^T \\ y^T H_1 \end{bmatrix} - \begin{bmatrix} 0, & 0 \\ y^T D, & y^T H_1 Y \end{bmatrix} \overline{N}^{-1} \begin{bmatrix} D^T \\ Y^T H_1 \end{bmatrix} \right),
\end{aligned}
$$

which is exactly the relation ($*$) with the matrix N determined by (ND) (Remark 46).

Theorem 33 Let H_1 be a SPD matrix and let for an arbitrary index $1 \le k \le m$ the formula (H) hold, where the matrix N_k is defined by the relation (NB) (the BFGS method). Then we can write

$$H_{k+1} = H_1 - [D_k, H_1 Y_k] \begin{bmatrix} 0, & R_k \\ R_k^T, & C_k + Y_k^T H_1 Y_k \end{bmatrix}^{-1} \begin{bmatrix} D_k^T \\ Y_k^T H_1 \end{bmatrix} \qquad (\overline{\mathrm{HB}})$$

Proof For $k = 1$, ($\overline{\mathrm{HB}}$) is equivalent to (HB) (i.e. to (H) where the matrix N is determined by (NB)). The induction step: Suppose that ($\overline{\mathrm{HB}}$) holds for all indices less than k. At the stage k, the formula ($\overline{\mathrm{HB}}$) can be writen in the form

$$H_+ = H_1 - [D, H_1 Y, d, H_1 y] \begin{bmatrix} 0, & R, & 0, & D^T y \\ R^T, & C + Y^T H_1 Y, & 0, & Y^T H_1 y \\ 0, & 0, & 0, & d^T y \\ y^T D, & y^T H_1 Y, & d^T y, & d^T y + y^T H_1 y \end{bmatrix}^{-1} \begin{bmatrix} D^T \\ Y^T H_1 \\ d^T \\ y^T H_1 \end{bmatrix}$$

(after permutations which simplify the notation). Using Lemma 11 and denoting

$$\overline{N} = \begin{bmatrix} 0, & R \\ R^T, & C + Y^T H_1 Y \end{bmatrix}$$

we obtain

$$\begin{aligned} H_+ &= H_1 - [D, H_1 Y] \overline{N}^{-1} \begin{bmatrix} D^T \\ Y^T H_1 \end{bmatrix} - \\ &\quad \left([d, H_1 y] - [D, H_1 Y] \overline{N}^{-1} \begin{bmatrix} 0, & D^T y \\ 0, & Y^T H_1 y \end{bmatrix} \right) \cdot \\ &\quad \left(\begin{bmatrix} 0, & d^T y \\ d^T y, & d^T y + y^T H_1 y \end{bmatrix} - \begin{bmatrix} 0, & 0 \\ y^T D, & y^T H_1 Y \end{bmatrix} \overline{N}^{-1} \begin{bmatrix} 0, & D^T y \\ 0, & Y^T H_1 y \end{bmatrix} \right)^{-1} \cdot \\ &\quad \left(\begin{bmatrix} d^T \\ y^T H_1 \end{bmatrix} - \begin{bmatrix} 0, & 0 \\ y^T D, & y^T H_1 Y \end{bmatrix} \overline{N}^{-1} \begin{bmatrix} D^T \\ Y^T H_1 \end{bmatrix} \right) = \\ &= H - \left([d, H_1 y] - [D, H_1 Y] \overline{N}^{-1} \begin{bmatrix} 0, & D^T y \\ 0, & Y^T H_1 y \end{bmatrix} \right) \cdot \\ &\quad \begin{bmatrix} 0, & d^T y \\ d^T y, & d^T y + y^T H y \end{bmatrix}^{-1} \left(\begin{bmatrix} d^T \\ y^T H_1 \end{bmatrix} - \begin{bmatrix} 0, & 0 \\ y^T D, & y^T H_1 Y \end{bmatrix} \overline{N}^{-1} \begin{bmatrix} D^T \\ Y^T H_1 \end{bmatrix} \right), \end{aligned}$$

which is exactly the relation $(*)$ with the matrix N determined by (NB) (Remark 46).

Theorem 34 Let H_1 be a SPD matrix and let for an arbitrary index $1 \le k \le m$ holds

$$H_{k+1} = H_k + (d_k - H_k y_k)(d_k^T y_k - y_k^T H_k y_k)^{-1} (d_k - H_k y_k)^T \qquad (\mathrm{HR})$$

(the symmetric rank 1 method). Then we can write

$$H_{k+1} = H_1 + (D_k - H_1 Y_k)(R_k + R_k^T - C_k - Y_k^T H_1 Y_k)^{-1} (D_k - H_1 Y_k)^T. \qquad (\overline{\mathrm{HR}})$$

Proof For $k = 1$, ($\overline{\mathrm{HR}}$) is equivalent to (HR). The induction step: Suppose that ($\overline{\mathrm{HR}}$) holds for all indices less than k. At the stage k, the formula ($\overline{\mathrm{HR}}$) can be written in the form

$$H_+ = H_1 - [D - H_1 Y, d - H_1 y] \begin{bmatrix} R + R^T - C - Y^T H_1 Y, & D^T y - Y^T H_1 y \\ y^T D - y^T H_1 Y, & d^T y - y^T H_1 y \end{bmatrix}^{-1} \begin{bmatrix} D^T - Y^T H_1 \\ d^T - y^T H_1 \end{bmatrix}.$$

Using Lemma 11 and denoting

$$\overline{N} = R + R^T - C - Y^T H_1 Y$$

we obtain

$$\begin{aligned}
H_+ &= H_1 + (D - H_1 Y)\overline{N}^{-1}(D - H_1 Y)^T + \\
&\left[(D - H_1 Y)\overline{N}^{-1}(D - H_1 Y)^T y - (d - H_1 y)\right] \cdot \\
&\left(d^T y - y^T H_1 y - y^T (D - H_1 Y)\overline{N}^{-1}(D - H_1 Y)^T y\right)^{-1} \cdot \\
&\left[(D - H_1 Y)\overline{N}^{-1}(D - H_1 Y)^T y - (d - H_1 y)\right]^T = \\
&= H + (d - Hy)\left(d^T y - y^T H y\right)^{-1}(d - Hy)^T
\end{aligned}$$

which is exactly the relation (HR).

Remark 47 Similar compact formulas can be derived for the matrix $B = H^{-1}$. We can use duality for this purpose. (Remark 27). Since the replacement $D_k \to Y_k$, $Y_k \to D_k$ is necessary, the upper half part of the matrix $D_k^T Y_k$ has to be replaced by the upper half part of the matrix $Y_k^T D_k$ or by the transposed lower half part of the matrix $D_k^T Y_k$. Therefore, instead of the upper triangular matrix R_k, we use the lower triangular matrix L_k such that $(L_k)_{ij} = 0$, $i < j$, and $(L_k)_{ij} = d_i^T y_j$, $i \geq j$. For the DFP method we obtain

$$B_{k+1} = B_1 - [Y_k, B_1 D_k]\begin{bmatrix} 0, & L_k^T \\ L_k, & C_k + D_k^T B_1 D_k \end{bmatrix}^{-1}\begin{bmatrix} Y_k^T \\ D_k^T B_1 \end{bmatrix}. \tag{$\overline{\mathrm{BD}}$}$$

For the BFGS method we obtain

$$B_{k+1} = B_1 - [Y_k, B_1 D_k]\begin{bmatrix} -C_k, & (L_k - C_k)^T \\ L_k - C_k, & D_k^T B_1 D_k \end{bmatrix}^{-1}\begin{bmatrix} Y_k^T \\ D_k^T B_1 \end{bmatrix}. \tag{$\overline{\mathrm{BB}}$}$$

For the symmetric rank 1 method we obtain

$$B_{k+1} = B_1 + (Y_k - B_1 D_k)\left(L_k + L_k^T - C_k - D_k^T B_1 D_k\right)^{-1}(Y_k - B_1 D_k)^T \qquad (\overline{\mathrm{BR}}).$$

Now we show how the compact formulas can be used in connection with limited storage variable metric methods. We restrict our attention to the limited storage BFGS method, which is in general the most efficient update among the updates described above. Using the substitution $H_1 = \gamma_k I$, where $\gamma_k = d_k^T y_k / y_k^T y_k$, the matrix ($\overline{\mathrm{HB}}$) can be expressed in the form

$$H_{k+1} = \gamma_k I + [D_k, \gamma_k Y_k]\begin{bmatrix} (R_k^{-1})^T (C_k + \gamma_k Y_k^T Y_k) R_k^{-1}, & -(R_k^{-1})^T \\ -R_k^{-1}, & 0 \end{bmatrix}\begin{bmatrix} D_k^T \\ \gamma_k Y_k^T \end{bmatrix}$$

(it can be proved using explicit inversion as in the proof of Lemma 11). Now we set $D_k = [d_{k-m}, \ldots, d_{k-1}]$ and $Y_k = [y_{k-m}, \ldots, y_{k-1}]$. The matrix C_k contains the diagonal of the matrix $D_k^T Y_k$ and the matrix R_k contains the upper half part of the matrix $D_k^T Y_k$. The matrices D_{k+1}, Y_{k+1} are obtained from the matrices D_k, Y_k easily by deleting the first and adding the last column. Similarly easily are obtained the matrices $D_{k+1}^T Y_{k+1}$, $Y_{k+1}^T Y_{k+1}$ from the matrices $D_k^T Y_k$, $Y_k^T Y_k$ and, therefore, the matrices C_{k+1}, R_{k+1} from the matrices C_k, R_k. Thus are available all matrices which are necessary for the computation of the matrix H_{k+1}.

Using the substitution $H_1 = \gamma_k I$, where $\gamma_k = d_k^T y_k / y_k^T y_k$, the matrix ($\overline{\mathrm{BB}}$) can be expressed in the form

$$\begin{aligned}
B_{k+1} &= \frac{1}{\gamma_k} I - \left[Y_k, \frac{1}{\gamma_k} D_k\right]\begin{bmatrix} -C_k^{-\frac{1}{2}}, & (\overline{L}_k^{-1}(L_k - C_k))^T \\ 0, & (\overline{L}_k^{-1})^T \end{bmatrix} \cdot \\
&\begin{bmatrix} C_k^{-\frac{1}{2}}, & 0 \\ -\overline{L}_k^{-1}(L_k - C_k), & \overline{L}_k^{-1} \end{bmatrix}\begin{bmatrix} Y_k^T \\ \frac{1}{\gamma_k} D_k^T \end{bmatrix},
\end{aligned}$$

where

$$\overline{L}_k \overline{L}_k^T = (L_k - C_k)^T C_k^{-\frac{1}{2}} (L_k - C_k) + \frac{1}{\gamma_k} D_k^T D_k$$

(it can be proved using explicit inversion and multiplication). We can see that only a matrix of order k has to be decomposed (for obtaining the matrix $\overline{L}_k \overline{L}_k^T$). While the expression ($\overline{\mathrm{HB}}$) is not much more advantageous than the Strang formula (R1) and (R2), the relation ($\overline{\mathrm{BB}}$) is very useful, since it can be utilized when it is necessary to work with the matrix B.

4.2 Difference version of the inexact Newton method

The difference version of the inexact Newton method is actually the inexact trust region method (Algorithm 5), where the matrix $B = G$ is not used but the multiplication $q = Bp = Gp$ is replaced by numerical differentiation

$$G(x)p \approx \frac{g(x + \delta p) - g(x)}{\delta},$$

where δ is a small difference ($\delta = \sqrt{\varepsilon_M} / \parallel p \parallel$ and ε_M is a machine precision). All other parts of Algorithm 5 remain unchanged. If the calculation of the gradient vector requires $O(n)$ operations, then this approach can be more efficient than matrix by vector multiplication ($O(n^2)$ operations in general). Moreover, the Hessian matrix has to be neither computed nor stored.

4.3 Sparse discrete Newton method

The discrete Newton method is based on the approximation of columns Ge_i, $1 \le i \le n$, of the Hessian matrix G by differences of gradients

$$G(x)e_i \approx \frac{g(x + \delta e_i) - g(x)}{\delta}, \qquad 1 \le i \le n,$$

where δ is a small difference ($\varepsilon = \sqrt{\varepsilon_M}$). However, if the Hessian matrix G is sparse, then usually one difference of gradients allows us to determine several columns of G. As a demonstrative example, we use the following band matrix

$$G = \begin{bmatrix} G_{11}, & G_{12}, & 0, & 0, & 0 \\ G_{21}, & G_{22}, & G_{23}, & 0, & 0 \\ 0, & G_{32}, & G_{33}, & G_{34}, & 0 \\ 0, & 0, & G_{43}, & G_{44}, & G_{45} \\ 0, & 0, & 0, & G_{54}, & G_{55} \end{bmatrix} \qquad \text{(G1)}.$$

If

$$\begin{aligned} v_1 &= [1, 0, 0, 1, 0]^T, \\ v_2 &= [0, 1, 0, 0, 1]^T, \\ v_3 &= [0, 0, 1, 0, 0]^T, \end{aligned}$$

then

$$\begin{aligned} Gv_1 &= [G_{11}, G_{21}, G_{34}, G_{44}, G_{54}]^T, \\ Gv_2 &= [G_{12}, G_{22}, G_{32}, G_{45}, G_{55}]^T, \\ Gv_3 &= [0, G_{23}, G_{33}, G_{43}, 0]^T, \end{aligned}$$

so that all elements of the matrix G are determined by three differences of gradients

$$
\begin{aligned}
\frac{g(x+\delta v_1)-g(x)}{\delta} &\approx Gv_1, \\
\frac{g(x+\delta v_2)-g(x)}{\delta} &\approx Gv_2, \\
\frac{g(x+\delta v_3)-g(x)}{\delta} &\approx Gv_3.
\end{aligned}
$$

The approach used in this particular case can be easily generalized. Divide columns of the matrix G into k disjoint groups $\mathcal{S}_i \subset \{1,\ldots,n\}$, $1 \le i \le k$, so that each submatrix $G(\mathcal{S}_i)$, consisting of columns of the matrix G belonging to the group $\mathcal{S}_i$, has at most one nonzero element in each row. Then all nonzero elements of the matrix G can be determined using k differences

$$
\frac{g(x+\delta v_i)-g(x)}{\delta} \approx Gv_i, \qquad 1 \le i \le k,
$$

where v_i, $1 \le i \le k$, are vectors, consisting only of zeroes and units, such that

$$
(v_i)_j = e_j^T v_i = 1 \Longleftrightarrow j \in \mathcal{S}_i
$$

(using the vector v_i we determine elements of the submatrix $G(\mathcal{S}_i)$). This way can be applied to an arbitrary (also nonsymmetric) matrix G. If the matrix G is symmetric, then symmetry can be used for an additional decrease of the number of gradient differences. Consider the matrix

$$
G = \left[\begin{array}{ccccc}
G_{11}, & G_{12}, & G_{13}, & G_{14}, & G_{15} \\
G_{21}, & G_{22}, & 0, & 0, & 0 \\
G_{31}, & 0, & G_{33}, & 0, & 0 \\
G_{41}, & 0, & 0, & G_{44}, & 0 \\
G_{51}, & 0, & 0, & 0, & G_{55}
\end{array}\right]. \tag{G2}
$$

If we use the above way, we require 5 gradient differences for determination of all nonzero elements of the matrix G. However, if we set

$$
\begin{aligned}
v_1 &= [1,0,0,0,0]^T, \\
v_2 &= [0,1,1,1,1]^T,
\end{aligned}
$$

then

$$
\begin{aligned}
Gv_1 &= [G_{11}, G_{21}, G_{31}, G_{41}, G_{51}]^T, \\
Gv_2 &= [*, G_{22}, G_{33}, G_{44}, G_{55}]^T,
\end{aligned}
$$

where the asterisk denotes the element, which is not interesting. Therefore, we have determined the elements G_{11}, G_{21}, G_{31}, G_{41}, G_{51}, G_{22}, G_{33}, G_{44}, G_{55} and, since the matrix G is symmetric, also the elements $G_{12} = G_{21}$, $G_{13} = G_{31}$, $G_{14} = G_{41}$, $G_{15} = G_{51}$, just by using two gradient differences.

The way which was used in this particular case can be again generalized. We again divide columns of the matrix G into k disjoint groups $\mathcal{S}_i$, $1 \le i \le k$. However, for determination of these groups, we do not use the original matrix G, but only its submatrices, obtained by deleting known rows and columns. Let G_i be the submatrix of G, obtained by deleting rows and columns with the indices $j \in S_1 \cup \ldots \cup S_{i-1}$, and let $G_i(\mathcal{S}_i)$ be the submatrix of G_i, containing only columns with the indices $j \in \mathcal{S}_i$, so that $G_i(\mathcal{S}_i) = G_i \cap G(\mathcal{S}_i)$.

If we divide columns of the matrix G into k disjoint groups $\mathcal{S}_i$, $i \in [1,k]$, so that each submatrix $G_i(\mathcal{S}_i)$ has at most one nonzero element in each row, then all nonzero elements of the matrix G can be determined using k differences

$$
\frac{g(x+\delta v_i)-g(x)}{\delta} \approx Gv_i, \qquad 1 \le i \le k,
$$

where v_i, $1 \le i \le k$, are vectors consisting only of zeroes and units, such that

$$(v_i)_j = e_j^T v_i = 1 \Longleftrightarrow j \in \mathcal{S}_i$$

(using the vector v_i we determine elements of the submatrix $G_i(\mathcal{S}_i) = G_i \cap G(\mathcal{S}_i)$; remaining elements of the submatrix $G(\mathcal{S}_i)$ can be determined from the symmetry of G).

We have not considered the determination of the groups $\mathcal{S}_i$, $1 \le i \le$. It is advantageous to choose these groups in such a way that its number be minimal. However, this is a complex combinatorial problem, which is equivalent to a certain graph coloring problem. In practice, simple but sufficiently fast algorithms are used, which find sufficiently small (not minimal in general) number of groups. The sequential process is usually used for the determination of groups $\mathcal{S}_i$, $1 \le i \le k$. Columns of the submatrix G_i are first ordered by a suitable rule and then they are used successively by increasing indices. The index $j \in \{1, \ldots, n\} \backslash (S_1 \cup \ldots \cup S_{i-1})$ is added to the group $\mathcal{S}_i$ only if the requirement, that the submatrix $G_i(\mathcal{S}_i)$ has at most one nonzero element in each row, is not violated.

Column ordering of the submatrix G_i is usually essential. The following matrices differ only by row and column orderings (nonzero elements are denoted by $*$).

$$\begin{bmatrix} * & & & & * \\ & * & & & * \\ & & * & & * \\ & & & * & * \\ * & * & * & * & * \end{bmatrix} \qquad \begin{bmatrix} * & * & * & * & * \\ * & * & & & \\ * & & * & & \\ * & & & * & \\ * & & & & * \end{bmatrix}$$

Using columns of the first matrix successively by increasing indices, we require 5 gradient differences for the determination of all nonzero elements. Using columns of the second matrix successively by increasing indices, we require only 2 gradient differences for the same purpose.

Besides the above direct methods for numerical computation of elements of the Hessian matrix, there exist indirect methods based on suitable substitutions, which usually require less number of gradient differences. Consider again the matrix (G1) and set

$$\begin{aligned} v_1 &= [1, 0, 1, 0, 1]^T, \\ v_2 &= [0, 1, 0, 1, 0]^T. \end{aligned}$$

Then

$$\frac{g(x + \delta v_1) - g(x)}{\delta} \approx G v_1 = \begin{bmatrix} G_{11} \\ G_{21} + G_{23} \\ G_{33} \\ G_{43} + G_{45} \\ G_{55} \end{bmatrix},$$

$$\frac{g(x + \delta v_2) - g(x)}{\delta} \approx G v_2 = \begin{bmatrix} G_{12} \\ G_{22} \\ G_{32} + G_{34} \\ G_{44} \\ G_{54} \end{bmatrix}.$$

Using these equations, we directly determine the values G_{11}, G_{33}, G_{55}, G_{12}, G_{22}, G_{44}, G_{54} and, since the matrix G is symmetric, the values G_{21}, G_{45}. Substituting the values G_{21}, G_{45} back into the above equations, we obtain the values G_{23}, G_{43} and, since the matrix G is symmetric, the values G_{32}, G_{34}. Only 2 gradient differences are necessary in this case (direct method requires 3 gradient differences).

The way, which was used in this particular case, can be easily generalized. Let G_U be the upper triangular matrix whose upper half part has the same sparsity pattern as upper half part of the matrix G. Divide columns of the matrix G_U into k disjoint groups $\mathcal{S}_i \subset \{1, \ldots, n\}$, $1 \le i \le k$, so that each submatrix $G_U(\mathcal{S}_i)$, consisiting of columns of the matrix G_U belonging to the group $\mathcal{S}_i$, has at most one nonzero element in each row. Then all nonzero elements of the matrix G can be determined using k differences

$$\frac{g(x+\delta v_i)-g(x)}{\delta} \approx G v_i, \qquad 1 \le i \le k$$

where v_i, $1 \le i \le k$, are vectors, consisting only of zeroes and units, such that

$$(v_i)_j = e_j^T v_i = 1 \Longleftrightarrow j \in \mathcal{S}_i$$

(using the vector v_i we determine elements of the submatrix $G_U(\mathcal{S}_i)$; remaining elements of the submatrix $G(\mathcal{S}_i)$ can be determined from the symmetry of G).

The main purpose of the above considerations was to show that the numerical computation of the Hessian matrix can be sufficiently cheap in the sparse case. This assigns a new role to the discrete Newton method, which can be competetive with variable metric methods or conjugate gradient methods in the sparse case.

Sparse discrete Newton method is usually realized as a trust region method with the optimal locally constrained step (Algorithm 4) or as a inexact trust region method (Algorithm 5). Trust region method with the optimal locally constrained step requires repeated solution of the system of linear equations $(G+\lambda I)s+g=0$ (for various values of the parameter $\lambda \ge 0$). The sparse Choleski decomposition

$$R^T R = P(G+\lambda I)P^T$$

is usually used, where R is a nonsingular upper triangular matrix and P is a permutation matrix, whose main purpose is the ordering of rows and columns of the matrix $G+\lambda I$ in such a way to minimize the fill-in. Determination of the permutation matrix P and the sparsity pattern of the upper triangular matrix R (symbolic decomposition) is carried out only once (before the iterative process is started), so that very sophisticated but complex methods can be used. Computation of elements of the upper triangular matrix R (numerical decomposition) is carried out by formulas given in Section 3.6.

Inexact trust region method is based on the conjugate gradient method used for solution of the linear system $Gs+g=0$. The sparse Hessian matrix G is used only for computation of products $q_i = Gp_i$, $1 \le i \le n$, so that it may not be decomposed. In the sparse case, it is advantageous to use a preconditioned conjugate gradient method which uses the recurrence relations

$$s_1 = 0, \qquad g_1 = g \qquad p_1 = -C^{-1} g$$

and

$$q_i = Bp_i, \qquad \alpha_i = g_i^T C^{-1} g_i / p_i^T q_i, \tag{$\overline{\mathrm{CG}}$}$$

$$s_{i+1} = s_i + \alpha_i p_i,$$

$$g_{i+1} = g_i + \alpha_i q_i, \qquad \beta_i = g_{i+1}^T C^{-1} g_{i+1} / g_i^T C^{-1} g_i,$$

$$p_{i+1} = -C^{-1} g_{i+1} + \beta_i p_i.$$

These recurrence relations can be obtained if we apply the basic conjugate gradient method (CG) to the minimization of the quadratic function

$$\tilde{Q}(\tilde{s}) = \tilde{g}^T \tilde{s} + \frac{1}{2} \tilde{s}^T \tilde{B} \tilde{s},$$

where $\tilde{g} = C^{-\frac{1}{2}} g$, $\tilde{s} = C^{\frac{1}{2}} s$, $\tilde{B} = C^{-\frac{1}{2}} B C^{-\frac{1}{2}}$ and if we set $g_i = C^{\frac{1}{2}} \tilde{g}_i$, $s_i = C^{-\frac{1}{2}} \tilde{s}_i$, $p_i = C^{-\frac{1}{2}} \tilde{p}_i$, $q_i = C^{\frac{1}{2}} \tilde{q}_i$, $1 \le i \le n$ (formulas for $\tilde{g}_i$ and $\tilde{q}_i$ have to be multiplied by the matrix $C^{\frac{1}{2}}$ and formulas for $\tilde{s}_i$ a $\tilde{p}_i$ have to be multiplied by the matrix $C^{-\frac{1}{2}}$).

Significance of preconditioning consists in that the matrix C is chosen to make the matrix $\tilde{B}$ better conditioned in comparison with the matrix B. Thus the method ($\overline{\mathrm{CG}}$) converges faster then the method (CG). An ideal choice is $C = B$. In this case, the minimum of the quadratic function $\tilde{Q}$ is reached already in the first step.

In connection with the sparse discrete Newton method, the preconditioning by the incomplete Choleski decomposition was shown to be very efficient. An idea of this way consists in using the Choleski decomposition formulas but omitting all new nonzero elements (fill-in). The resulting upper triangular matrix R has the same sparsity pattern as the upper half part of the matrix B and the approximation $RR^T \approx B$ is usually satisfactory, which leads to very efficient preconditioning.

Now we describe a way for the representation of the sparse Hessian matrix G in the computer storage. We denote by G_U the upper triangular matrix obtained from the matrix G by putting to zero all subdiagonal elements.

Definition 25 Sparse representation of the Hessian matrix G is a triple of vectors $num\ (G_U) \in R^{m_U}$, $ind\ (G_U) \in R^{m_U}$, $ord\ (G_U) \in R^{n+1}$, where m_U is the number of nonzero elements of the matrix G_U. The vector $num\ (G_U)$ contains numerical values of nonzero elements of the matrix G_U sorted by rows. The vector $ind\ (G_U)$ contains column indices of these nonzero elements. The vector $ord\ (G_U)$ contains pointers of diagonal elements of the matrix G_U (pointers are related to the vectors $num\ (G_U)$ and $ind\ (G_U)$). The last element of the vector $ord\ (G_U)$ (with the index $n+1$) has the value m_U+1.

The matrix (G1) is represented in the form

$$\begin{aligned} num(G_U) &= [G_{11}, G_{12}, G_{22}, G_{23}, G_{33}, G_{34}, G_{44}, G_{45}, G_{55}]^T, \\ ind(G_U) &= [1, 2, 2, 3, 3, 4, 4, 5, 5]^T, \\ ord(G_U) &= [1, 3, 5, 7, 9, 10]^T. \end{aligned}$$

For the matrix (G2) it holds

$$\begin{aligned} num(G_U) &= [G_{11}, G_{12}, G_{13}, G_{14}, G_{15}, G_{22}, G_{33}, G_{44}, G_{55}]^T, \\ ind(G_U) &= [1, 2, 3, 4, 5, 2, 3, 4, 5]^T, \\ ord(G_U) &= [1, 6, 7, 8, 9, 10]^T. \end{aligned}$$

4.4 Sparse variable metric methods

Sparse variable metric methods are based on updates which preserve the sparsity pattern of the Hessian matrix. Preserving sparsity pattern is a strong restriction, which eliminates some important properties of variable metric methods. Updates cannot have a low rank in general. For instance, a diagonal update of a diagonal matrix, which changes it to satisfy the quasi-Newton condition, can have rank n in general. Moreover, positive definiteness of the updated matrix can be lost for an arbitrary sparse update, which can be again demonstrated on a diagonal matrix. From this point of view, it is interesting that a superlinear rate of convergence can be obtained even if the quadratic termination property does not hold.

Sparse variable metric updates should satisfy the quasi-Newton condition, not violate symmetry and preserve sparsity. Denote

$$\begin{aligned} \mathcal{V}_Q &= \{B \in R^{n \times n} : Bd = y\} \\ \mathcal{V}_S &= \{B \in R^{n \times n} : B^T = B\} \\ \mathcal{V}_G &= \{B \in R^{n \times n} : G_{ij} = 0 \Rightarrow B_{ij} = 0\} \end{aligned}$$

(we assume, that $G_{ii} \neq 0\ \forall 1 \leq i \leq n$). Clearly, $\mathcal{V}_Q \subset R^{n \times n}$, $\mathcal{V}_S \subset R^{n \times n}$, $\mathcal{V}_G \subset R^{n \times n}$ are linear manifolds ($\mathcal{V}_S$ and $\mathcal{V}_G$ are subspaces) in $R^{n \times n}$. Since the Frobenius norm is an Euclidean norm in $R^{n \times n}$, we can define orthogonal projections $\mathcal{P}_Q$, $\mathcal{P}_S$, $\mathcal{P}_G$ into $\mathcal{V}_Q$, $\mathcal{V}_S$, $\mathcal{V}_G$ by the formulas

$$
\begin{aligned}
\mathcal{P}_Q B &= \arg \min_{B^+ \in \mathcal{V}_Q} \| B^+ - B \|_F, \\
\mathcal{P}_S B &= \arg \min_{B^+ \in \mathcal{V}_S} \| B^+ - B \|_F, \\
\mathcal{P}_G B &= \arg \min_{B^+ \in \mathcal{V}_G} \| B^+ - B \|_F \, .
\end{aligned}
$$

Similarly, we can define orthogonal projections $\mathcal{P}_{QS}$, $\mathcal{P}_{QG}$, $\mathcal{P}_{SG}$ and $\mathcal{P}_{QSG}$ into $\mathcal{V}_Q \cap \mathcal{V}_S$, $\mathcal{V}_Q \cap \mathcal{V}_G$, $\mathcal{V}_S \cap \mathcal{V}_G$ and $\mathcal{V}_Q \cap \mathcal{V}_S \cap \mathcal{V}_G$. It is clear, that the requirements laid down on a sparse update are satisfed by the matrix $B^+ = \mathcal{P}_{QSG} B$.

Theorem 35 Let $B \in R^{n \times n}$ and let $\mathcal{P}_Q$, $\mathcal{P}_S$, $\mathcal{P}_G$ be orthogonal projections into $\mathcal{V}_Q$, $\mathcal{V}_S$, $\mathcal{V}_G$. Then

$$
\begin{aligned}
\mathcal{P}_Q B &= B + \frac{(y - Bd)d^T}{d^T d}, \\
\mathcal{P}_S B &= \frac{1}{2}(B + B^T)
\end{aligned}
$$

and

$$
\begin{aligned}
(\mathcal{P}_G B)_{ij} &= B_{ij} \ , \ G_{ij} \neq 0, \\
(\mathcal{P}_G B)_{ij} &= 0 \quad \ , \ G_{ij} = 0.
\end{aligned}
$$

Proof We proceed similarly as in the proof of Theorem 19. The formula for $\mathcal{P}_Q B$ can be derived using the Lagrangian function

$$
\begin{aligned}
L &= \frac{1}{2} \| B^+ - B \|_F^2 + u^T (w - (B^+ - B)d) = \\
&= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (B_{ij}^+ - B_{ij})^2 + \sum_{i=1}^{n} u_i \left(y_i - \sum_{j=1}^{n} B_{ij}^+ d_j \right) .
\end{aligned}
$$

Differentiating this Lagrangian function with respect to elements of the matrix B^+, we obtain

$$
\frac{\partial L}{\partial B_{ij}^+} = (B_{ij}^+ - B_{ij}) - u_i d_j
$$

$\forall 1 \leq i \leq n$, $\forall 1 \leq j \leq n$. Therefore, the stationarity condition for the Lagrangian function has the form $B^+ = B + ud^T$, which after substitution into the quasi-Newton condition $B^+ d = y$ gives $ud^T d = y - Bd$, or

$$
u = \frac{y - Bd}{d^T d}.
$$

Substituting this expression into the formula $B^+ = B + ud^T$, we obtain the relation for $\mathcal{P}_Q B$. The relation for $\mathcal{P}_S B$ can be derived by minimization of the function

$$
\frac{1}{2} \| B^+ - B \|_F^2 = \frac{1}{2} \sum_{i=1}^{n} (B_{ii}^+ - B_{ii})^2 + \frac{1}{2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} ((B_{ij}^+ - B_{ij})^2 + (B_{ij}^+ - B_{ji})^2),
$$

Differentiating this function with respect to elements of the matrix B^+ and setting the first order derivatives equal to zero, we obtain the conditions

$$
\begin{aligned}
B_{ij}^+ - B_{ij} &= 0 \ , \ i = j, \\
(B_{ij}^+ - B_{ij}) + (B_{ij}^+ - B_{ji}) &= 0 \ , \ i, \neq j
\end{aligned}
$$

which give $B^+ = (B + B^T)/2$. The relation for $\mathcal{P}_G B$ can be derived by minimization of the function

$$\frac{1}{2} \parallel B^+ - B \parallel_F^2 = \frac{1}{2} \sum_{G_{ij} \neq 0} (B_{ij}^+ - B_{ij})^2 + \frac{1}{2} \sum_{G_{ij}=0} B_{ij}^2,$$

since by the assumption $B_{ij}^+ = 0$ if $G_{ij} = 0$. Differentiating this function with respect to elements of the matrix B^+ and setting the first order derivatives equal to zero, we obtain

$$\begin{aligned} B_{ij}^+ - B_{ij} &= 0 \ , \ G_{ij} \neq 0, \\ B_{ij}^+ &= 0 \ , \ G_{ij} = 0, \end{aligned}$$

which was necessary to prove.

To eliminate the zero elements from the quasi-Newton condition, we define vectors $d^i \in R^n$, $1 \le i \le n$, in such a way that

$$\begin{aligned} d_j^i &= d_j \ , \ G_{ij} \neq 0, \\ d_j^i &= 0 \ \ , \ G_{ij} = 0 \end{aligned}$$

and instead of the standard quasi-Newton condition $B^+ d = y$ we use the sparse quasi-Newton condition

$$\sum_{j=1}^{n} (B_{ij}^+ - (\mathcal{P}_G B)_{ij}) d_j^i = w_i, \qquad 1 \le i \le n,$$

where $w = y - (\mathcal{P}_G B)d$.

Theorem 36 Let $B \in R^{n \times n}$ and let $\mathcal{P}_{QS}, \mathcal{P}_{QG}, \mathcal{P}_{SG}$ be orthogonal projections into $\mathcal{V}_Q \cap \mathcal{V}_S$, $\mathcal{V}_Q \cap \mathcal{V}_G$, $\mathcal{V}_S \cap \mathcal{V}_G$. Then

$$\begin{aligned} \mathcal{P}_{QS} B &= B + \frac{(y - Bd)d^T + d(y - Bd)^T}{d^T d} - \frac{(y - Bd)^T d}{d^T d} \frac{dd^T}{d^T d}, \\ \mathcal{P}_{QG} B &= \mathcal{P}_G (B + ud^T), \\ \mathcal{P}_{SG} B &= \mathcal{P}_S \mathcal{P}_G B = \mathcal{P}_G \mathcal{P}_S B, \end{aligned}$$

where the vector $u \in R^n$ is a solution of the linear system $Qu = w$ with the diagonal positive semidefinite matrix

$$Q = \sum_{i=1}^{n} \parallel d^i \parallel^2 e_i e_i^T.$$

Proof The formula for $\mathcal{P}_{QS} B$ follows immediately from Theorem 19 (the PSB method). It suffices to substitute $W = I$. It is clear that $\mathcal{P}_{QG} B = \mathcal{P}_{QG} \mathcal{P}_G B$ holds. The relation for $\mathcal{P}_{QG} B = \mathcal{P}_{QG} \mathcal{P}_G B$ can be derived using the Lagrangian function

$$L = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (B_{ij}^+ - (\mathcal{P}_G B)_{ij})^2 + \sum_{i=1}^{n} u_i \left(w_i - \sum_{j=1}^{n} (B_{ij}^+ - (\mathcal{P}_G B)_{ij}) d_j^i \right)$$

containing the sparse quasi-Newton condition. Differentiating this Lagrangian function with respect to elements of the matrix B^+, we obtain

$$\frac{\partial L}{\partial B_{ij}^+} = (B_{ij}^+ - (\mathcal{P}_G B)_{ij}) - u_i d_j^i$$

$\forall 1 \leq i \leq n$, $\forall 1 \leq j \leq n$. Therefore, the stationarity condition for the Lagrangian function has the form $B^+_{ij} - (\mathcal{P}_G B)_{ij} = u_i d^i_j$, $1 \leq i \leq n$, $1 \leq j \leq n$. Substituting this relation into the sparse quasi-Newton condition, we obtain

$$\sum_{j=1}^{n} u_i d^i_j d^i_j = w_i, \qquad 1 \leq i \leq n$$

or $Qu = w$, where Q is a diagonal matrix appearing in statement of the theorem (positive semidefinitenss is obvious). The relations for $\mathcal{P}_{SG}B$ follow immediately from the identities $\mathcal{P}_{SG}B = \mathcal{P}_{SG}(\mathcal{P}_S B)$ and $\mathcal{P}_{SG}B = \mathcal{P}_{SG}(\mathcal{P}_G B)$.

Theorem 37 Let $B \in R^{n\times n}$ and let $\mathcal{P}_{QSG}$ be an orthogonal projection into $\mathcal{V}_Q \cap \mathcal{V}_S \cap \mathcal{V}_G$. Then

$$\mathcal{P}_{QSG}B = \mathcal{P}_G(B + ud^T + du^T),$$

where the vector $u \in R^n$ is a solution of the linear system $Qu = w$ with the symmetric positive semidefinite matrix

$$Q = \mathcal{P}_G(dd^T) + \sum_{i=1}^{n} \| d^i \|^2 e_i e_i^T,$$

which has the same sparsity pattern as the matrix G.

Proof It is clear that $\mathcal{P}_{QSG}B = \mathcal{P}_{QSG}\mathcal{P}_G B$ holds. Since the matrix $B^+ - \mathcal{P}_G B$ is symmetric, we can set $B^+ - \mathcal{P}_G B = X + X^T$, where X is a square matrix to be determined. We use the Lagrangian function

$$L = \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_{ij} + X_{ji})^2 + \sum_{i=1}^{n} u_i \left(w_i - \sum_{j=1}^{n}(X_{ij} + X_{ji})d^i_j \right)$$

containing the sparse quasi-Newton condition. Differentiating this Lagrangian function with respect to the elements of the matrix X, we obtain

$$\frac{\partial L}{\partial X_{ij}} = (X_{ij} + X_{ji}) - u_i d^i_j - u_j d^j_i$$

$\forall 1 \leq i \leq n$, $\forall 1 \leq j \leq n$. Therefore, the stationarity condition for the Lagrangian function has the form $B^+ - (\mathcal{P}_G B)_{ij} = u_i d^i_j + u_j d^j_i$. Substituting this expression into the sparse quasi-Newton condition, we obtain

$$w_i = \sum_{j=1}^{n}(u_i d^i_j + u_j d^j_i)d^i_j = \| d^i \|^2 u_i + \sum_{j=1}^{n} d^j_i d^i_j u_j$$

or $Qu = w$, where Q is a symmetric matrix appearing in the statement of the theorem. Obviously, the matrix Q has the same sparsity pattern as the matrix G. Let $v \in R^n$ be an arbitrary vector. Then

$$\begin{aligned} v^T Q v &= \sum_{i=1}^{n}\sum_{j=1}^{n} d^j_i d^i_j v_i v_j + \sum_{i=1}^{n} \| d^i \|^2 v_i^2 = \sum_{G_{ij}\neq 0} d_i d_j v_i v_j + \sum_{G_{ij}\neq 0} d_j^2 v_i^2 = \\ &= \frac{1}{2}\sum_{G_{ij}\neq 0} (d_i v_j + d_j v_i)^2 \geq 0 \end{aligned} \tag{$*$}$$

(the matrix G_{ij} is symmetric), so that the matrix Q is positive semidefinite. It remains to prove that the equation $Qu = w$ has a solution. Suppose first that $\| d^i \| \neq \forall 1 \leq i \leq n$. We show that the matrix

Q is positive definite in this case. If it is not true, then a vector $v \neq 0$ exists, such that $v^T Qv = 0$. Then, by expression $(*)$, an index $1 \le i \le n$ exists, such that $v_i \neq 0$ and

$$d_i v_j + d_j v_i = 0 \qquad \forall G_{ij} \neq 0.$$

Since we suppose that $G_{ii} \neq 0$, then necessarily $d_i v_i = 0$, or $d_i = 0$, which after substitution into the last equality gives $d_j v_i = 0 \; \forall G_{ij} \neq 0$, or $d_j = 0 \; \forall G_{ij} \neq 0$, which is a contradiction to the assumption that

$$\| d^i \|^2 = \sum_{j=1}^{n} (d^i_j)^2 = \sum_{G_{ij} \neq 0} d_j^2 \neq 0.$$

Suppose now that for some index $1 \le i \le n$ $\| d^i \| = 0$ holds. Then the matrix Q has only zeros in the i-th row and i-th column and

$$w_i = y_i - \sum_{G_{ij} \neq 0} B_{ij} d_j = \sum_{G_{ij} \neq 0} (\tilde{G}_{ij} - B_{ij}) d^i_j = 0$$

holds (the matrix $\tilde{G}$ is defined by relation (a) in the proof of Lemma 6). Therefore, we can omit the i-th equation and set $u_i = 0$. In this way, we eliminate all unnecessary equations. The remained system has a positive definite matrix.

The sparse variable metric method, which uses the update

$$B^+ = \mathcal{P}_{QSG} B \tag{BT}$$

is called the Toint method. Its implementation is relatively difficult, since an additional linear system $Qu = w$ has to be solved (the Toint system). In the dense case, this method is equivalent to the PSB method, which is not too efficient. Therefore, easier updates with better convergence properties were looked for. Such updates usually violate the sparse quasi-Newton condition. In this section, we consider the Marwil update

$$B^+ = \mathcal{P}_S \mathcal{P}_{QG} B \qquad \text{(BM)},$$

the Powell update

$$B^+ = \mathcal{P}_G \mathcal{P}_{QS} B \tag{BP}$$

and the Steihaug update

$$B^+ = \mathcal{P}_{SG} \mathcal{P}_Q B \qquad \text{(BS)}.$$

Lemma 12 Let B^+ be a matrix determined by one of the updates (BT), (BM), (BP), (BS). Then

$$B^+ \in \mathcal{V}_S \cap \mathcal{V}_G.$$

Proof For the updates (BT) and (BS), the statement is obvious. For the update (BM), the statement follows from the fact that the projection $\mathcal{P}_S$, determined by symmetry of the matrix, does not violate a symmetric sparsity pattern. For the update (BP), the statement follows from the fact that the projection $\mathcal{P}_G$, determined by a symmetric sparsity pattern, does not violate symmetry.

In the formulas (BT), (BM), (BP), (BS), always two orthogonal projections $\mathcal{P}_A$, $\mathcal{P}_B$ into the linear manifolds $\mathcal{V}_A$, $\mathcal{V}_B$ occur, such that $\mathcal{V}_A \subset \mathcal{V}_Q$ and $\mathcal{V}_A \cap \mathcal{V}_B = \mathcal{V}_Q \cap \mathcal{V}_S \cap \mathcal{V}_G$ (the second projection is an identity in the case of the Toint update).

Lemma 13 Let $B^+ = \mathcal{P}_B\mathcal{P}_A B$, where $\mathcal{P}_A, \mathcal{P}_B$ are orthogonal projections into $\mathcal{V}_A$, $\mathcal{V}_B$, where $\mathcal{V}_A \subset R^{n\times n}$, $\mathcal{V}_B \subset R^{n\times n}$ are linear manifolds such that $\mathcal{V}_A \subset \mathcal{V}_Q$ and $\mathcal{V}_A \cap \mathcal{V}_B = \mathcal{V}_Q \cap \mathcal{V}_S \cap \mathcal{V}_G$. Then for an arbitrary matrix $\tilde{G} \in \mathcal{V}_Q \cap \mathcal{V}_S \cap \mathcal{V}_G$ it holds

$$\| B^+ - \tilde{G} \|_F^2 \le \| B - \tilde{G} \|_F^2 - \frac{\| y - Bd \|^2}{\| d \|^2}.$$

Proof Since $\tilde{G} \in \mathcal{V}_B$ and $\mathcal{P}_B$ is an orthogonal projection, we can use Pythagoras theorem

$$\| \mathcal{P}_B\mathcal{P}_A B - \tilde{G} \|_F^2 = \| \mathcal{P}_A B - \tilde{G} \|_F^2 - \| \mathcal{P}_B\mathcal{P}_A B - \mathcal{P}_A B \|_F^2 \le \| \mathcal{P}_A B - \tilde{G} \|_F^2 .$$

Since $\mathcal{P}_A B \in \mathcal{V}_A \subset \mathcal{V}_Q$, we can write $\mathcal{P}_A Bd = y$, so that

$$\| y - Bd \| = \| (\mathcal{P}_A B - B)d \| \le \| \mathcal{P}_A B - B \| \| d \| \le \| \mathcal{P}_A B - B \|_F \| d \| .$$

Since $\tilde{G} \in \mathcal{V}_A$ and $\mathcal{P}_A$ is an orthogonal projection, we can write

$$\| \mathcal{P}_A B - \tilde{G} \|_F^2 = \| B - \tilde{G} \|_F^2 - \| \mathcal{P}_A B - B \|_F^2 .$$

Using all above inequalities we obtain the statement of the lemma.

Now we will investigate the convergence of sparse variable metric methods. We restrict ourselves to the trust region methods, since sparse variable metric updates do not guarantee positive definiteness of updated matrices in general.

Theorem 38 Let $x_i \in R^n$, $i \in N$, be a sequence of points qenerated by a trust region method (Definition 22). Let $B_{i+1} = \mathcal{P}_B\mathcal{P}_A(B_i)$, $i \in N_2$, let $B_{i+1} = B_i$, $i \notin N_2$ (here $\mathcal{P}_B\mathcal{P}_A(B_i)$ denotes one of the sparse updates (BT), (BM), (BP), (BS) and the sets N_1, N_2, N_3 be defined as in Remark 38). If the function $F : R^n \to R$ satisfies the conditions (F1), (F2), (F3) and (F5), then

$$\liminf_{i\to\infty} \| g_i \| = 0.$$

Proof (a) first we prove that $\| B_i \| \le C_i$, $i \in N$, where C_i, $i \in N$, are numbers satisfying the recurrence inequalities

$$C_{i+1} \le C_i + \overline{C} \| d_i \|, \qquad i \in N. \tag{C}$$

Let $i \in N_2$ and let $\tilde{G}_i$ be the matrix defined by the relation (a) in the proof of Lemma 6. Then

$$\begin{aligned} \| \tilde{G}_i - G_i \|_F &= \| \int_0^1 (G(x_i + \lambda d_i) - G(x_i))d\lambda \|_F \le \sqrt{n} \int_0^1 \| G(x_i + \lambda d_i) - G(x_i) \| d\lambda \le \\ &\le \overline{L}\sqrt{n} \| d_i \| \int_0^1 \lambda d\lambda = \frac{1}{2}\overline{L}\sqrt{n} \| d_i \| \end{aligned}$$

(we use the assumption (F5) and the fact that the Frobenius norm is not greater than $\sqrt{n}$ times the spectral norm). In the similar way we obtain

$$\| \tilde{G}_i - G_{i+1} \|_F \le \frac{1}{2}\overline{L}\sqrt{n} \| d_i \| .$$

If we use the inequality $\| B_{i+1} - \tilde{G}_i \|_F \le \| B_i - \tilde{G}_i \|_F$, which follows from Lemma 13, we can write

$$\begin{aligned} \| B_{i+1} - G_{i+1} \|_F &\le \| B_{i+1} - \tilde{G}_i \|_F + \| \tilde{G}_i - G_{i+1} \|_F \le \| B_i - \tilde{G}_i \|_F + \| \tilde{G}_i - G_{i+1} \|_F \le \\ &\le \| B_i - G_i \|_F + \| \tilde{G}_i - G_i \|_F + \| \tilde{G}_i - G_{i+1} \|_F \le \\ &\le \| B_i - G_i \|_F + \overline{L}\sqrt{n} \| d_i \| . \end{aligned}$$

This inequality is satisfied even for $i \notin N_2$, since $i \notin N_2$ implies $B_{i+1} = B_i$, $G_{i+1} = G_i$ and $d_i = 0$, and using it successively we obtain

$$\| B_{i+1} - G_i \|_F \leq \| B_1 - G_1 \|_F + \overline{L}\sqrt{n} \sum_{j=1}^{i} \| d_j \| .$$

Using the assumption (F3) and setting $C_1 = 2\overline{G}\sqrt{n} + \| B_1 \|_F$ and $\overline{C} = \overline{L}\sqrt{n}$, we get the inequalities $\| B_i \| \leq C_i$, $i \in N$, where the numbers C_i, $i \in N$, satisfy the inequalities (C).
(b) From the assumption (F2), the existence of a constant $\overline{\Delta}$ follows, such that $\| d_i \| \leq \overline{\Delta}$, $i \in N$. Choose this constant sufficiently great such that $C_1 \leq \overline{C\Delta}$ holds. Since $\| d_i \| \leq \overline{\Delta}$, $i \in N$, and since (C) holds, we can write $C_i \leq i\overline{C\Delta}$, $i \in N$. Denoting

$$M_i = \max_{1 \leq j \leq i} \| B_j \|,$$

we can write $M_i \leq C_i$, $i \in N$, so that

$$\sum_{i=1}^{\infty} \frac{1}{M_i} \geq \sum_{i=1}^{\infty} \frac{1}{C_i} \geq \frac{1}{\overline{C\Delta}} \sum_{i=1}^{\infty} \frac{1}{i} = \infty$$

holds and we can use Theorem 21.

Proposition 8 Let the assumptions of Theorem 38 be satisfied and let $x_i \to x^*$. Let the function $F : R^n \to R$, satisfy the conditions (F3), (F4), (F5). Then

$$\sum_{i=1}^{\infty} \| d_i \| < \infty.$$

Remark 48 This proposition, whose proof is formally rather complicated, implies the existence of a constant $\overline{B}$, such that $\| B_i \| \leq \overline{B}$ $\forall i \in N$, since from (C) it easily follows

$$\| B_i \| \leq C_i \leq C_1 + \overline{C} \sum_{j=1}^{i-1} \| d_i \| \leq C_1 + \overline{C} \sum_{j=1}^{\infty} \| d_i \| \triangleq \overline{B}.$$

Theorem 39 Let the assumptions of Proposition 8, be satisfied, where $\| \omega_i(s_i) \| \to 0$. Then $x_i \to x^*$ Q-superlinearly.

Proof Let $i \in N_2$. Using Lemma 13 and the first two inequalities from the proof of Theorem 38, we can write

$$\begin{aligned} \| B_{i+1} - G_{i+1} \|_F^2 &\leq \left(\| B_{i+1} - \tilde{G}_i \|_F + \| G_{i+1} - \tilde{G}_i \|_F \right)^2 \leq \\ &\leq \| B_{i+1} - \tilde{G}_i \|_F^2 + \frac{1}{4}\overline{L}^2 n \| d_i \|^2 + \overline{L}\sqrt{n} \| B_{i+1} - \tilde{G}_i \|_F \| d_i \| \leq \\ &\leq \| B_i - \tilde{G}_i \|_F^2 - \frac{\| y_i - B_i d_i \|^2}{\| d_i \|} + \left(\frac{1}{4}\overline{L}^2 n\overline{\Delta} + \overline{L}\sqrt{n}\,(\overline{B} + \overline{G}) \right) \| d_i \| \end{aligned}$$

and

$$\begin{aligned} \| B_i - \tilde{G}_i \|_F^2 &\leq \left(\| B_i - G_i \|_F + \| G_i - \tilde{G}_i \|_F \right)^2 \leq \\ &\leq \| B_i - G_i \|_F^2 + \frac{1}{4}\overline{L}^2 n \| d_i \|^2 + \overline{L}\sqrt{n} \| B_i - G_i \|_F \| d_i \| \leq \\ &\leq \| B_i - G_i \|_F^2 + \left(\frac{1}{4}\overline{L}^2 n\overline{\Delta} + \overline{L}\sqrt{n}\,(\overline{B} + \overline{G}) \right) \| d_i \| \end{aligned}$$

(the existence of a constant $\overline{\Delta}$ follows from Proposition 8, the existence of a constant $\overline{B}$ follows from Remark 48, the existence of a constant $\overline{G}$ follows from (F3)). Using both these inequalities, we get

$$\frac{\| y_i - B_i d_i \|^2}{\| d_i \|^2} \leq \| B_i - G_i \|_F^2 - \| B_{i+1} - G_{i+1} \|_F^2 + \overline{M} \| d_i \|,$$

where $\overline{M} = \frac{1}{2}\overline{L}^2 n\overline{\Delta} + 2\overline{L}\sqrt{n}(\overline{B} + \overline{G})$. This inequality holds formally also for $i \notin N_2$ (if $i \notin N_2$, then $d_i = 0$, $y_i = 0$ and the relation on the left hand side can be replaced by zero). Using this inequality together with Proposition 8, we obtain

$$\begin{aligned}\sum_{i=1}^{\infty} \frac{\| y_i - B_i d_i \|^2}{\| d_i \|^2} &\leq \sum_{i=1}^{\infty} (\| B_i - G_i \|_F^2 - \| B_{i+1} - G_{i+1} \|_F^2) + \overline{M} \sum_{i=1}^{\infty} \| d_i \| \leq \\ &\leq \| B_1 - G_1 \|_F^2 + \overline{M} \sum_{i=1}^{\infty} \| d_i \| < \infty.\end{aligned} \tag{$*$}$$

Furthermore, (F5) implies

$$\begin{aligned}\frac{\| (G^* - B_i) d_i \|}{\| d_i \|} &\leq \frac{\| G^* d_i - y_i \|}{\| d_i \|} + \frac{\| y_i - B_i d_i \|}{\| d_i \|} \leq \\ &\leq \| G^* - G_i \| + \| \tilde{G}_i - G_i \| + \frac{\| y_i - B_i d_i \|}{\| d_i \|} \leq \\ &\leq \overline{L} \| x^* - x_i \| + \frac{1}{2}\overline{L}\sqrt{n} \| d_i \| + \frac{\| y_i - B_i d_i \|}{\| d_i \|}\end{aligned}$$

(see proof of Theorem 38), so that

$$\frac{\| (G^* - B_i) d_i \|}{\| d_i \|} \to 0,$$

since $x_i \to x^*$ by the assumption, $\| d_i \| \to 0$ by Proposition 8 and $\| y_i - B_i d_i \| / \| d_i \| \to 0$ by $(*)$. Since $\| \omega_i(s_i) \| \to 0$, the assumptions of Theorem 22 are satisfied and $x_i \to x^*$ Q-superlinearly.

Sparse variable metric methods can be also realized in the line search framework, where an inexact solution of the linear system $Bs + g = 0$ is obtained by the conjugate gradient method (Section 3.4).

Theorem 40 Applying the conjugate gradient method (CG) to the quadratic function $Q(s) = (1/2)s^T Bs + g^T s$, where $\underline{B} \leq \underline{\lambda}(B) \leq \overline{\lambda}(B) \leq \overline{B}$, and assuming that $p_i^T B p_i > 0$, $1 \leq i \leq m$, we obtain the vector $s = s_{m+1}$ satisfying the condition

$$-s^T g \geq \varepsilon_0 \| s \| \| g \|$$

with

$$\varepsilon_0 = \frac{1}{m} \frac{\underline{B}}{\overline{B}}.$$

Proof From the proof of Theorem 25 it follows that

$$\begin{aligned} p_j^T B p_i &= 0 \qquad \forall 1 \leq j < i \leq m, \\ g_j^T g_i &= 0 \qquad \forall 1 \leq j < i \leq m+1, \\ p_j^T g_i &= 0 \qquad \forall 1 \leq j < i \leq m+1. \end{aligned}$$

Using (CG), we obtain

$$g_i^T p_i = \left(g + \sum_{j=1}^{i-1} \alpha_j B p_j \right)^T p_i = g^T p_i$$

and

$$\alpha_i = \frac{g_i^T g_i}{p_i^T B p_i} = -\frac{g_i^T (p_i - \beta_{i-1} p_{i-1})}{p_i^T B p_i} = -\frac{g^T p_i}{p_i^T B p_i}$$

$\forall 1 \leq i \leq m$, so that

$$-g^T s = -\sum_{i=1}^{m} \alpha_i g^T p_i = \sum_{i=1}^{m} \frac{(g^T p_i)^2}{p_i^T B p_i} \geq \frac{(g^T p_1)^2}{p_1^T B p_1} = \frac{g^T g}{g^T B g} \| g \|^2 \geq \| g \|^2 / \overline{B}.$$

Furthermore,

$$s = \sum_{i=1}^{m} \alpha_i p_i = -\sum_{i=1}^{m} \frac{p_i p_i^T}{p_i^T B p_i} g,$$

so that

$$\| s \| \leq \sum_{i=1}^{m} \frac{\| p_i p_i^T \|}{p_i^T B p_i} \| g \| = \sum_{i=1}^{m} \frac{p_i^T p_i}{p_i^T B p_i} \| g \| \leq m \| g \| / \underline{B}.$$

Using both these inequalities, we obtain

$$-\frac{s^T g}{\| s \| \| g \|} \geq \frac{\| g \|^2}{\overline{B}} \frac{\underline{B}}{m \| g \|^2} = \frac{1}{m} \frac{\underline{B}}{\overline{B}}.$$

Statement of Theorem 40 is a basis for the following simple algorithm:

Algorithm 7 Data $0 < \omega < 1$, $0 < \overline{\varepsilon} < 1$, $m \geq n$.

Step 1 Set $s = 0$, $r = -g$, $\sigma = \| r \|^2$, $p = r$, $\delta = \| r \|^2$, $k = 1$.

Step 2 Set $\rho = \sigma$, compute the vector $q = Bp$ and the numbers $\tau = p^T q$, $\gamma = \tau / \delta$. If $k = 1$, then set $\overline{\gamma} = \gamma$. If $k = 1$ and $\gamma \leq 0$, then set $s = r$ and terminate the computation. If $k > 1$ and $\gamma \leq \overline{\varepsilon}\overline{\gamma}$, then terminate the computation.

Step 3 Set $\alpha = \rho / \tau$. Set $s := s + \alpha p$, $r := r - \alpha q$ and $\sigma = \| r \|^2$. If either $\sigma \leq \omega^2 \| g \|^2$ or $k \geq m$, then terminate the computation.

Step 4 Set $\beta = \sigma / \rho$, $p = r + \beta p$, $\delta := \sigma + \beta^2 \delta$, $k := k + 1$ and go to Step 2.

We usually choose $m = n + 3$.

The use of the conjugate gradient method is very advantageous, since this method, applied to the quadratic function $Q(s)$ with a positive definite matrix B, generates descent directions independently of the accuracy of the solution to the linear system $Bs + g$ (the general Theorem 4 requires only that $\| Bs + g \| \leq \overline{\omega} \| g \|$ and $\overline{\omega} \leq \underline{B}/\overline{B}$) hold. Even if the convergence theory, which we have investigated, cannot be applied to a line search realizations of sparse variable metric methods, these realizations are usually more efficient than trust region realizations.

The following table shows a comparison of several methods for sparse problems of unconstrained optimization (CG - conjugate gradient method, 5-BFGS - 5-step limited storage BFGS method, difference version of the inexact Newton method, sparse Newton method with either iterative CG or direct GM solution of the linear system, sparse variable metric (Marwil) method with either iterative CG or direct GM solution of the linear system, dense BFGS method) by means of 10 test problems with 100 variables (the total number of iterations NIT, function evaluations NFV and gradient evaluations NFG, are given together with the total computational time).

Method	NIT - NFV - NFG	time
CG	2066 - 3903 - 3903	7.19
5-BFGS	1965 - 2149 - 2149	7.63
Inexact Newton + CG	702 - 822 - 6188	7.25
Sparse Newton + CG	518 - 567 - 2461	7.69
Sparse Newton + GM	377 - 405 - 1722	7.03
Sparse Marwil + CG	1318 - 2009 - 2009	12.58
Sparse Marwil + GM	2440 - 4841 - 4841	36.41
Dense BFGS	1498 - 1656 - 1656	23.73

4.5 Partitioned discrete Newton method

Large-scale problems are frequently formulated in such a way that

$$F(x) = \sum_{k=1}^{m} f_k(x),$$

where $m = O(n)$ and where each one of the functions $f_k : R^n \to R$, $1 \le k \le m$, depends only on $n_k = O(1)$ variables. Then the evaluation of the value and the gradient of the objective function $F(x)$ needs $O(n)$ operations and the Hessian matrix contains $O(n)$ nonzero elements. The gradient and the Hessian matrix of the function $F : R^n \to R$ can be expressed in the form

$$g(x) = \sum_{k=1}^{m} g_k(x),$$

$$G(x) = \sum_{k=1}^{m} G_k(x),$$

where the gradients $g_k(x)$ and the Hessian matrices $G_k(x)$ of the functions $f_k : R^n \to R$, $1 \le k \le m$, contain $O(1)$ nonzero elements, so that they can be stored in a packed form. Denote

$$\begin{aligned} f(x) &= [f_1(x), \ldots, f_m(x)]^T, \\ J(x) &= [g_1(x), \ldots, g_m(x)]^T. \end{aligned}$$

Then $F(x) = f^T(x)e$, $g(x) = J^T(x)e$ holds, where $e = [1, \ldots, 1]^T \in R^m$ is a vector containing only unit elements. The Jacobian matrix $J(x)$ is sparse (its k-th row $g_k^T(x)$ contains $n_k = O(1)$ nonzero elements, $1 \le k \le m$). The Hessian matrix $G(x)$ has the same sparsity pattern ast the matrix $J^T(x)J(x)$. Therefore, the sparsity pattern of the Hessian matrix is fully determined by the sparsity pattern of the Jacobian matrix.

Definition 26 A sparse representation of the sparse Jacobian matrix J is defined as a triple of vectors $num\ (J) \in R^{\hat{n}}$, $ind\ (J) \in R^{\hat{n}}$, $ord\ (J) \in R^{m+1}$, where

$$\hat{n} = \sum_{k=1}^{m} n_k$$

is the number of nonzero elements of the matrix J. The vector $num\ (J)$ contains numerical values of nonzero elements of the matrix J sorted by rows. The vector $ind\ (J)$ contains column indices of these nonzero elements. The vector $ord\ (J)$ contains pointers of the first elements in the rows of the matrix J (pointers are related to the vectors $num\ (J)$ and $ind\ (J)$), so that

$$ord(J)_k = 1 + \sum_{i=1}^{k-1} n_k, \qquad 1 \le k \le m + 1.$$

In the sequel, we will use the packed gradients $\hat{g}_k(x) \in R^{n_k}$ containing only nonzero elements of the gradients $g_k(x) \in R^n$, $1 \le k \le m$, and packed Hessian matrices $\hat{G}_k(x) \in R^{n_k \times n_k}$ containing only nonzero elements of the Hessian matrices $G_k(x) \in R^{n \times n}$, $1 \le k \le m$. It is clear that

$$num(J) = [\hat{g}_1^T, \dots, \hat{g}_m^T]^T.$$

The partitioned discrete Newton method is based on the numerical computation of elements of the packed Hessian matrices. The following difference formulas are used

$$\hat{G}_k(x)\hat{e}_j \approx \frac{\hat{g}_k(x + \delta \hat{e}_j) - \hat{g}_k(x)}{\delta},$$

where $\hat{e}_j$, $1 \le j \le n_k$, are columns of the unit matrix of order n_k. Thus the determination of all elements of the packed Hessian matrices requires

$$\sum_{k=1}^{m} n_k^2 = mO(1) = O(n)$$

operations.

Remark 49 The sparse Jacobian matrix J and the numerical values of the packed Hessian matrices $\hat{G}_k(x)$, $1 \le k \le m$, uniquely determine the gradient g and the sparse Hessian matrix G. Knowing the sparse representation of the Jacobian matrix (Definition 26) and the numerical values of the packed Hessian matrices, we can easily determine the sparse representation of the Hessian matrix (Definition 25). The packed Hessian matrices can be used successively (it is not necessary to store them simultaneously).

The partitioned discrete Newton method differs from the sparse discrete Newton method only in the way for obtaining the sparse Hessian matrix $G(x)$. All other considerations remain unchanged. We can again use a trust region realization with optimal locally constrained step (Section 3.3) or inexact trust region realization (Section 3.4).

Numerical comparison of the partitioned and the sparse discrete Newton methods shows that both these types require approximately the same number of operations per iteration and require approximately the same number of iterations. The sparse discrete Newton method is algorithmically more complicated (it is necessary to find a suitable partition of columns of the Hessian matrix) but all complicated operations are carried out only once, before the iterative process is started. Therefore, the total computational time may be even slightly shorter.

4.6 Partitioned variable metric methods

Partitioned variable metric methods use approximations $\hat{B}_k$, $1 \le k \le m$, of the packed Hessian matrices $\hat{G}_k(x)$, $1 \le k \le m$, which are constructed using the variable metric updates

$$\hat{B}_k^+ = \frac{1}{\hat{\gamma}_k}\left(\hat{B}_k + \frac{\hat{\gamma}_k}{\hat{\rho}_k}\frac{1}{\hat{b}_k}\hat{y}_k\hat{y}_k^T - \frac{1}{\hat{c}_k}\hat{B}_k\hat{d}_k\left(\hat{B}_k\hat{d}_k\right)^T + \frac{\hat{\beta}_k}{\hat{c}_k}\left(\frac{\hat{c}_k}{\hat{b}_k}\hat{y}_k - \hat{B}_k\hat{d}_k\right)\left(\frac{\hat{c}_k}{\hat{b}_k}\hat{y}_k - \hat{B}_k\hat{d}_k\right)^T\right),$$

where $\hat{y}_k = \hat{g}_k^+ - \hat{g}_k$ and $\hat{d}_k \in R^{n_k}$ is a vector, which has the same dimension as the vector $\hat{y}_k$ and which contains elements of the vector d with indices corresponding to elements of the vector $\hat{y}_k$ (all for $1 \le k \le m$). Moreover, $\hat{b}_k = \hat{y}_k^T\hat{d}_k$, $\hat{c}_k = \hat{d}_k^T\hat{B}_k\hat{d}_k$, $1 \le k \le m$, and $\hat{\gamma}_k$, $\hat{\rho}_k$, $\hat{\beta}_k$, $1 \le k \le m$ are free parameters.

First we state several comments concerning partitioned variable metric methods:

- Partitioned variable metric methods are usually more efficient than sparse variable metric methods, which follows from the numerical comparison given at the end of this section.
- Since the packed matrices $\hat{B}_k$, $1 \le k \le m$, are updated using the vectors y_k, d_k, $1 \le k \le m$, it is advantageous that $\hat{B}_k \to \hat{G}_k$, $1 \le k \le m$ to holds, so that we usually set $\gamma_k = 1$, $\rho_k = 1$, $1 \le k \le m$.

- It can be proved that the partitioned variable metric methods are Q-superlinearly convergent. Surprisingly, it is more dificult to prove their global convergence, which was not done without additional complicated and restrictive conditions. This is due to the fact that the inequalities $\hat{y}_k^T \hat{d}_k > 0$, $1 \le k \le n$, may not be satisfied in general so that the matrices $\hat{B}_k^+$, $1 \le k \le m$, may not be positive definite.

Now we describe an efficient line search realization of partitioned variable metric methods. This realization uses the safeguarded BFGS updates

$$\begin{aligned} \hat{B}_k^+ &= \hat{B}_k + \frac{1}{\hat{b}_k}\hat{y}_k\hat{y}_k^T - \frac{1}{\hat{c}_k}\hat{B}_k\hat{d}_k\left(\hat{B}_k\hat{d}_k\right)^T, && \hat{y}_k^T\hat{d}_k > 0, \\ \hat{B}_k^+ &= \hat{B}_k &&, \quad \hat{y}_k^T\hat{d}_k \le 0 \end{aligned}$$

where $1 \le k \le m$. The above formulas guarantee positive definiteness of the matrices $\hat{B}_k^+$, $1 \le k \le m$. If the matrix $\hat{B}_k$ is positive definite and $\hat{y}_k^T\hat{d}_k \le 0$ holds, then the matrix $\hat{B}_k^+ = \hat{B}_k$ is obviously positive definite. If the matrix $\hat{B}_k$ is positive definite and $\hat{y}_k^T\hat{d}_k > 0$ holds, then the matrix $\hat{B}_k^+$ is positive definite by Theorem 16.

Knowing the approximations $\hat{B}_k$, $1 \le k \le m$ of the packed Hessian matrices $\hat{G}_k$, $1 \le k \le m$, we can determine the sparse representation of the approximation B of the sparse Hessian matrix G (Remark 49). However, partitioned variable metric methods have a disadvantage, which consists in the fact that the matrices $\hat{B}_k$, $1 \le k \le m$, have to be simultaneously stored. It requires additional

$$\hat{m} = \sum_{k=1}^{n} \frac{1}{2}\hat{n}_k(\hat{n_k} + 1)$$

storage elements (the number $\hat{m}$ is usually much greater than the number of nonzero elements of the sparse Hessian matrix G).

The following table shows a comparison of several methods for partitioned problems of unconstrained optimization (CG - conjugate gradient method, 5-BFGS - 5-step limited storage BFGS method, difference version of the inexact Newton method, partitioned Newton method with either iterative CG or direct GM solution of the linear system, partitioned variable metric (BFGS) method with either iterative CG or direct GM solution of the linear system) by means of 10 test problems with 100 variables (the total number of iterations NIT, function evaluations NFV and gradient evaluations NFG, are given together with the total computational time).

Method	NIT - NFV - NFG	time
CG	2390 - 4501 - 4501	12.75
5-BFGS	2389 - 2586 - 2586	11.91
Inexact Newton	644 - 813 - 6280	11.09
Partitioned Newton + CG	556 - 607 - 2328	16.64
Partitioned Newton + GM	416 - 446 - 1635	14.28
Partitioned BFGS + CG	841 - 995 - 995	12.96
Partitioned BFGS + GM	753 - 882 - 882	10.54

4.7 Sparse modifications of the Gauss-Newton method for nonlinear least squares

Suppose that the objective function $F(x)$ has the form

$$F(x) = \frac{1}{2}\sum_{k=1}^{m} f_k^2(x),$$

where $m = O(n)$ and where each one of the functions $f_k : R^n \to R$, $1 \le k \le m$, depends on $n_k = O(1)$ variables. Thus we have a special case of the partitioned problem. This partitioned problem can be solved by the partitioned discrete Newton method or by partitioned variable metric methods. However, the special form of the objective function allows us to use modifications of the Gauss-Newton method (hybrid methods), which can be much more efficient.

The Gauss-Newton (GN) method can be realized either by using the sparse Hessian representation (by solving the system of normal equations (NE)) or by using the sparse Jacobian representation (by solving the overdetermined system (OE)). The first way is based on the use of the matrix $B = J^T J$, which has the same sparsity pattern at the Hessian matrix G and which can be easily determined. Knowing the matrix B, we can realize the GN method either as a trust region method with an optimal locally constrained step or as an inexact trust region method (utilize the same considerations as in Section 4.3.)

Since the GN method can fail in the case of large residual problems, it is advantageous to combine this method with other methods as it is shown in Section 3.9. The following hybrid methods are very efficient in practical computations:

1) Combination of the GN method with the sparse Marwil update. In the first iteration we set $B = J^T J$. After finishing an iteration we set

$$\begin{aligned} B_+ &= J_+^T J_+ \quad , \quad F - F_+ > \underline{\vartheta} F, \\ B_+ &= \mathcal{P}_S \mathcal{P}_{QS} J_+^T J_+ \quad , \quad F - F_+ \le \underline{\vartheta} F \end{aligned}$$

(see (BM) in Section 4.4), where $J_+ = J(x_+)$. Global convergence of this method follows from Theorem 30 and Theorem 38. Superlinear convergence follows from Theorem 30 and Theorem 39.

2) Combination of the GN method with the discrete Newton method. In the first iteration we set $B = J^T J$. After finishing an iteration we set

$$\begin{aligned} B_+ &= J_+^T J_+ \quad , \quad F - F_+ > \underline{\vartheta} F, \\ B_+ &= J_+^T J_+ + \sum_{k=1}^{m} f_k^+ G_k^+ \quad , \quad F - F_+ \le \underline{\vartheta} F, \end{aligned}$$

where $J_+ = J(x_+)$ and $f_k^+ = f_k(x_+)$, $G_k^+ = G_k(x_+)$, $1 \le k \le m$ ($G_k(x_+)$ is a difference approximation of the Hessian matrix of the function $f_k(x_+)$). Global and superlinear convergence of this method follows from Theorem 29 and Theorem 30.

3) Combination of the GN method with the symmetric rank 1 (SR1) update. In the first iteration we set $B = J^T J$ and $B_k = I_k$, $1 \le k \le m$ (B_k is an approximation of the Hessian matrix G_k and I_k differs from the unit matrix only in that $(I_k)_{ii} = 0$ if $(G_k)_{ii} = 0$). After finishing an iteration we set

$$\begin{aligned} B_k^+ &= B_k + \frac{w_k w_k^T}{d_k^T w_k} \quad , \quad |d_k^T w_k| > \underline{\delta}, \\ B_k^+ &= B_k \quad , \quad |d_k^T w_k| \le \underline{\delta}, \end{aligned}$$

for $1 \le k \le m$, and

$$\begin{aligned} B_+ &= J_+^T J_+ \quad , \quad F - F_+ > \underline{\vartheta} F, \\ B_+ &= J_+^T J_+ + \sum_{k=1}^{m} f_k^+ B_k^+ \quad , \quad F - F_+ \le \underline{\vartheta} F, \end{aligned}$$

where $w_k = y_k - B_k d_k$ and $y_k = g_k(x_+) - g_k(x)$, $d_k = x_+ - x$, $1 \le k \le m$. Althouhg no satisfactory convergence proof is known, this method is very robust. A disadvantage of this method consist in the fact that the matrices B_k, $1 \le k \le m$ have to be simultaneously stored (in fact we use the packed matrices $\hat{B}_k$, $1 \le k \le m$).

The following table shows a comparison of several methods for nonlinear least squares, which use the sparse Hessian representation. Some general methods are also included. The table contains the total number of iterations NIT, function evaluations NFV and gradient evaluations NFG together with the number of failures and the total computational time. Results were obtained by means of 22 test problems with 100 variables.

Method	NIT - NFV - NFG	fail	time
GN	1584 -1819 - 1603	–	31.47
GN + sparse Marwil	1039 - 1136 - 1061	–	20.92
GN + partitioned SR1	980 - 1064 - 1002	–	28.24
GN + discrete Newton	947 - 1042 - 1247	–	25.92
Discrete Newton	1139 - 1269 - 3732	–	1:20.02
Sparse Marwil	3411 - 5131 - 5131	1	1:07.17
partitioned BFGS	3014 - 3793 - 3793	1	1:23.43
CG	8024 - 15670 - 15670	3	2:23.68
5 - BFGS	7106 - 7762 - 7762	3	1:47.00
Inexact Newton	1486 - 16608 - 16383	–	2:34.34

A failure means that 1000 iterations or 2000 function evaluations do not suffice for obtaining a solution. From the above table it follows that the sparse modifications of the GN method are more efficient than general methods for sparse or partitioned problems.

The sparse Hessian representation cannot be used if the Jacobian matrix contains at least one dense row ($n_k \sim n$ for some index $1 \le k \le m$). In this case, the matrix G is dense (the same sparsity pattern has the matrix J^TJ) so that we have to use the sparse Jacobian representation. If we utilize the sparse Jacobian approximation, then possibilities for including the second order information are limited (we do not consider this case). We focus our attention on iterative methods for solving the normal equations $J^TJs + J^Tf = 0$.

The most simple modification of the CG method for solving the normal equations is the CGNE method.

Definition 27 Let $J \in R^{m\times n}$ be a matrix with linearly independent columns and $f \in R^m$. Then the iterative process described by the recurrence relations

$$s_1 = 0, \qquad u_1 = f, \qquad g_1 = J^T f, \qquad p_1 = -g_1,$$

and

$$v_i = Jp_i, \qquad \alpha_i = \| g_i \|^2 \,/\, \| v_i \|^2,$$

$$s_{i+1} = s_i + \alpha_i p_i, \qquad u_{i+1} = u_i + \alpha_i v_i,$$

$$g_{i+1} = J^T u_{i+1}, \qquad \beta_i = \| g_{i+1} \|^2 \,/\, \| g_i \|^2,$$

$$p_{i+1} = -g_{i+1} + \beta_i p_i,$$

for $1 \le i \le n$, where $u_i \in R^m$, $v_i \in R^m$, $1 \le i \le n$, is the CGNE method defined by the matrix $J \in R^{m\times n}$ and the vector $f \in R^m$.

It can be easily verified (if we set $B = J^TJ$ and $q_i = J^Tv_i$, $1 \le i \le n$), That the CGNE method is equivalent to the CG method described in Section 3.4. Properties of the CGNE method are very similar to the properties of the CG method. However, if $m \gg n$, then the CGNE method requires more operations and more storage elements that the CG method.

The CGNE method can be replaced by methods based on the symmetric Lanczos process, which are more stable.

Definition 28 Let $B \in R^{n\times n}$ be a symmetric positive definite (SPD) matrix and $g \in R^n$. Then the iterative process described by the recurrence relations

$$q_0 = 0, \qquad \beta_1 q_1 = g$$

and

$$\alpha_i = q_i^T B q_i, \tag{SL}$$

$$\beta_{i+1} q_{i+1} = B q_i - \alpha_i q_i - \beta_i q_{i-1},$$

for $1 \le i \le n$, where the coefficients β_i, $1 \le i \le n$ are chosen for q_i, $1 \le i \le n$ to have unit norms, is the symmetric Lanczos process defined by the matrix $B \in R^{n\times n}$ and the vector $g \in R^n$.

Remark 50 Let $\beta_i \neq 0$, $1 \le i \le k$, for some index $1 \le k \le n$. Then (SL) implies $g = Q_k(\beta_1 e_1)$ and

$$B Q_k = Q_k T_k + \beta_{k+1} q_{k+1} e_k^T, \tag{$\overline{\text{SL}}$}$$

where $Q_k = [q_1, q_2, \dots, q_{k-1}, q_k]$, $e_1^T = [1, 0, \dots, 0, 0]$, $e_k^T = [0, 0, \dots, 0, 1]$ and

$$T_k = \begin{bmatrix} \alpha_1, & \beta_2, & \dots, & 0, & 0 \\ \beta_2, & \alpha_2, & \dots, & 0, & 0 \\ -- & -- & -- & -- & -- \\ 0, & 0, & \dots, & \alpha_{k-1}, & \beta_k \\ 0, & 0, & \dots, & \beta_k, & \alpha_k \end{bmatrix}$$

(the matrix $T_k \in R^{k\times k}$ is tridiagonal). It can be easily verified by the multiplication, using the relations (SL).

Theorem 41 Consider the symmetric Lanczos process defined by a SPD matrix $B \in R^{n\times n}$ and by a vector $g \in R^n$. Let $\beta_i \neq 0$, $1 \le i \le k$ for some index $1 \le k \le n$. Then the vectors q_i, $1 \le i \le k$, form an orthonormal basis of the Krylov subspace $\mathcal{K}_k = \text{span}\,(g, Bg, \dots, B^{k-1}g)$.

Proof (by induction). For $k = 1$, it is obvious since $q_1 = g/ \parallel g \parallel$. Assume that the statement holds for some index $1 \le k < n$ and that $\beta_{k+1} \neq 0$. Using the inductive assumption, we get $Q_k^T Q_k = I$, so that $Q_k^T B Q_k = T_k + \beta_{k+1} Q_k^T q_{k+1} e_k^T$. The matrix $Q_k^T B Q_k$ is symmetric and the same holds for the matrix T_k, so that necessarily $Q_{k-1}^T q_{k+1} = 0$. Furthermore, (SL) implies $\beta_{k+1} q_k^T q_{k+1} = q_k^T B q_k - \alpha_k = \alpha_k - \alpha_k = 0$. Hence the vector q_{k+1} is orthogonal to the vectors q_i, $1 \le i \le k$, and has unit norm. By (SL), the vectors q_i, $1 \le i \le k+1$ lie in the Krylov subspace $\mathcal{K}_{k+1}$ and since they are mutually orthogonal with unit norms, they form a basis there.

Remark 51 Since $Q_k^T Q_k = I$ and $Q_k^T q_{k+1} = 0$ (proof of Theorem 41), we can write

$$Q_k^T B Q_k = T_k,$$

so that the symmetric Lanczos process can be used for tridiagonalization of the matrix B.

Remark 52 The symmetric Lanczos process can be used for solution of the linear system $Bs + g = 0$. We set $s_1 = 0$ and construct the vectors s_{i+1}, $1 \le i \le k$, in such a way that

$$s_{i+1} = \arg\min_{s \in \mathcal{K}_i} \left(\frac{1}{2} s^T B s + g^T s \right)$$

holds. Since $s \in \mathcal{K}_i$ if and only if $s = Q_i z$, where $z \in R^i$, we can write $s_{i+1} = Q_i z_i$, where

$$z_i = \arg\min_{z \in R^i} \left(\frac{1}{2} z^T T_i z + \beta_1 e_1^T z \right)$$

(it follows from the relations $g = Q_i(\beta_1 e_1)$ and $Q_i^T Q_i = I$). If $\beta_{k+1} = 0$, then the vector $s_{k+1} \in \mathcal{K}_k$ is a solution of the linear system $Bs + g = 0$. Indeed, ($\overline{\text{SL}}$) implies $BQ_k = Q_k T_k$ and since the matrix T_k is nonsingular, we can set $z_k = -T_k^{-1}(\beta_1 e_1)$, which gives $B_k s_{k+1} = B Q_k z_k = -Q_k T_k T_k^{-1}(\beta_1 e_1) = -Q_k(\beta_1 e_1) = -g$.

Theorem 42 The vectors s_{i+1}, $1 \le i \le k$ defined in Remark 52 are identical with the vectors $\overline{s}_{i+1}$, $1 \le i \le k$, generated by the CG method which uses the following relations

$$\overline{s}_1 = 0, \quad \overline{g}_1 = g, \quad \overline{p}_1 = -\overline{g}_1$$

and

$$\overline{q}_i = B\overline{p}_i, \quad \overline{\alpha}_i = \| \overline{g}_i \|^2 / \overline{p}_i^T \overline{q}_i,$$

$$\overline{s}_{i+1} = \overline{s}_i + \overline{\alpha}_i \overline{p}_i,$$

$$\overline{g}_{i+1} = \overline{g}_i + \overline{\alpha}_i \overline{q}_i, \quad \overline{\beta}_i = \| \overline{g}_{i+1} \|^2 / \| \overline{g}_i \|^2,$$

$$\overline{p}_{i+1} = -\overline{g}_{i+1} + \overline{\beta}_i \overline{p}_i,$$

for $1 \le i \le k$. Moreover $\alpha_1 = 1/\overline{\alpha}_1$ and

$$\alpha_{i+1} = \frac{\overline{\beta}_i}{\overline{\alpha}_i} + \frac{1}{\overline{\alpha}_{i+1}},$$

$$\beta_{i+1} = \frac{\sqrt{\overline{\beta}_i}}{\overline{\alpha}_i},$$

$$q_i = (-1)^{i+1} \overline{g}_i / \| \overline{g}_i \|,$$

for $1 \le i \le k$.

Proof From proof of Theorem 10 we can deduce that the vectors $\overline{s}_{i+1}$, $1 \le i \le k$, determined by the CG method, lie in the Krylov subspaces $\mathcal{K}_i$, $1 \le i \le k$, and realize there the constrained minimima of the quadratic function $Q(s) = (1/2)s^T Bs + g^T s$. This is exactly a definition of the vectors s_{i+1}, $1 \le i \le k$, in Remark 52. Since the vectors $\overline{g}_i$, $1 \le i \le k$ are mutually orthogonal and lie in the Krylov subspaces $\mathcal{K}_i$, $1 \le i \le k$, they have to be collinear with the vectors q_i, $1 \le i \le k$, or

$$\overline{G}_k = Q_k D_k,$$

where $\overline{G}_k = [\overline{g}_1, \ldots, \overline{g}_k]$ and $D_k = \text{diag}\,(\varepsilon_1 \| \overline{g}_1 \|, \ldots, \varepsilon_k \| \overline{g}_k \|)$ (the numbers ε_i, $1 \le i \le k$, can obtain the values ± 1). Set $\overline{P}_k = [\overline{p}_1, \ldots, \overline{p}_k]$. Then recurrence relations of the CG method imply

$$\overline{G}_k = \overline{P}_k \overline{B}_k,$$

where

$$\overline{B}_k = \begin{bmatrix} -1, & \overline{\beta}_1, & \ldots, & 0 \\ 0, & -1, & \ldots, & 0 \\ -- & -- & -- & -- \\ 0, & 0, & \ldots, & -1 \end{bmatrix}$$

is an upper bidiagonal matrix. From proof of Theorem 10 it follows that the matrix $\overline{P}_k^T B \overline{P}_k$ is diagonal. Using recurrence relations of the CG method, we obtain $\overline{P}_k^T B \overline{P}_k = \text{diag}(\| \overline{g}_1 \|^2 / \overline{\alpha}_1, \ldots, \| \overline{g}_k \|^2 / \overline{\alpha}_k) = D_k \text{diag}(1/\overline{\alpha}_1, \ldots, 1/\overline{\alpha}_k) D_k$, so that

$$\begin{aligned} T_k &= Q_k^T B Q_k = D_k^{-1} \overline{G}_k^T B \overline{G}_k D_k^{-1} = D_k^{-1} \overline{B}_k^T \overline{P}_k^T B \overline{P}_k \overline{B}_k D_k^{-1} = \\ &= D_k^{-1} \overline{B}_k^T D_k \text{diag}\,(1/\overline{\alpha}_1, \ldots, 1/\overline{\alpha}_k) D_k \overline{B}_k D_k^{-1}. \end{aligned}$$

But

$$D_k B_k D_k^{-1} = \begin{bmatrix} -1, & \overline{\beta}_1 \frac{\varepsilon_1 \|\overline{g}_1\|}{\varepsilon_2 \|\overline{g}_2\|}, & \ldots, & 0 \\ 0, & -1, & \ldots, & 0 \\ -- & -- & -- & -- \\ 0, & 0, & \ldots, & -1 \end{bmatrix} = \begin{bmatrix} -1, & \varepsilon_1 \varepsilon_2 \sqrt{\overline{\beta}_1}, & \ldots, & 0 \\ 0, & -1, & \ldots, & 0 \\ -- & -- & -- & -- \\ 0, & 0, & \ldots, & -1 \end{bmatrix}.$$

Substituting this relation into the expression for the matrix T_k, we can write

$$T_k = \begin{bmatrix} \frac{1}{\overline{\alpha}_1}, & \frac{-\varepsilon_1\varepsilon_2\sqrt{\overline{\beta}_1}}{\overline{\alpha}_1}, & \ldots, & 0 \\ \frac{-\varepsilon_1\varepsilon_2\sqrt{\overline{\beta}_1}}{\overline{\alpha}_1}, & \frac{\overline{\beta}_1}{\overline{\alpha}_1} + \frac{1}{\overline{\alpha}_2}, & \ldots, & 0 \\ -- & -- & -- & -- \\ 0, & 0, & \ldots, & \frac{\overline{\beta}_{k-1}}{\overline{\alpha}_{k-1}} + \frac{1}{\overline{\alpha}_k} \end{bmatrix}.$$

which after comparison with $(\overline{\mathrm{SL}})$ gives $\alpha_1 = 1/\overline{\alpha}_1$ and

$$\begin{aligned} \alpha_{i+1} &= \frac{\overline{\beta}_i}{\overline{\alpha}_i} + \frac{1}{\overline{\alpha}_{i+1}}, \\ \beta_{i+1} &= -\frac{\varepsilon_i \varepsilon_{i+1} \sqrt{\overline{\beta}_i}}{\overline{\alpha}_i} = \frac{\sqrt{\overline{\beta}_i}}{\overline{\alpha}_i}, \end{aligned}$$

for $1 \le i \le k$ (since $\overline{\alpha}_i > 0$, $\overline{\beta}_i \ge 0$ $\beta_{i+1} \ge 0$, then $\varepsilon_i \varepsilon_{i+1} = -1$ has to hold). Sinse (SL) implies $\beta_1 q_1 = g = \overline{g}_1$ and $\beta_1 \ge 0$, we obtain $\varepsilon_1 = 1$, which together with $\varepsilon_1 \varepsilon_{i+1} = -1$, $1 \le i \le k$, gives $\varepsilon_i = (-1)^{i+1}$, so that

$$q_i = (-1)^{i+1} \overline{g}_i / \parallel \overline{g}_i \parallel .$$

Definition 29 Let $J \in R^{m \times n}$ be a matrix with linearly independent columns and $f \in R^m$. Then the iterative process described by the recurrence relations

$$\delta_1 u_1 = f, \qquad \gamma_1 q_1 = J^T u_1$$

and

$$\delta_{i+1} u_{i+1} = J q_i - \gamma_i u_i, \tag{BL}$$

$$\gamma_{i+1} q_{i+1} = J^T u_{i+1} - \delta_{i+1} q_i,$$

for $1 \le i \le n$, where the coefficients γ_i, δ_i, $1 \le i \le n$ are chosen for $u_i \in R^m$, $q_i \in R^n$, $1 \le i \le n$ to have unit norms, is the Lanczos bidiagonalization process defined by the matrix $J \in R^{m \times n}$ and the vector $f \in R^m$.

Remark 53 Let $\gamma_i \neq 0$, $\delta_i \neq 0$, $1 \le i \le k$ for some index $1 \le k \le n$. Then (BL) implies $f = U_{k+1}(\delta_1 e_1)$ and

$$J Q_k = U_{k+1} B_k,$$

$$J^T U_{k+1} = Q_k B_k^T + \gamma_{k+1} q_{k+1} e_{k+1}^T \tag{$\overline{\mathrm{BL}}$},$$

where $Q_k = [q_1, q_2, \ldots, q_k]$, $U_{k+1} = [u_1, u_2, \ldots, u_k, u_{k+1}]$, $e_1^T = [1, 0, \ldots, 0, 0]$, $e_{k+1}^T = [0, 0, \ldots, 0, 1]$ and

$$B_k = \begin{bmatrix} \gamma_1, & 0, & \ldots, & 0 \\ \delta_2, & \gamma_2, & \ldots, & 0 \\ -- & -- & -- & -- \\ 0, & 0, & \ldots, & \gamma_k \\ 0, & 0, & \ldots, & \delta_{k+1} \end{bmatrix}$$

(the matrix $B_k \in R^{(k+1)\times k}$ is bidiagonal).

Theorem 43 Consider the Lanczos bidiagonalization process defined by a matrix $J \in R^{m \times n}$ and a vector $f \in R^m$. Let $\gamma_i \neq 0$, $\delta_i \neq 0$, $1 \leq i \leq k$, for some index $1 \leq k \leq n$. Then the vectors q_i, $1 \leq i \leq k$, form an orthonormal basis in the Krylov subspace $\mathcal{K}_i = \text{span}\ (g, Bg, \dots, B^{k-1}g)$, where $B = J^T J$ and $g = J^T f$, and the vectors u_i, $1 \leq i \leq k$, are mutually ortogonal with unit norms.

Proof (by induction). For $k = 1$, it is obvious since $q_1 = J^T f / \parallel J^T f \parallel$ and $u_1 = f / \parallel f \parallel$. Suppose that the statement holds for some index $1 \leq k < n$ and that $\gamma_{k+1} \neq 0$, $\delta_{k+1} \neq 0$. Using ($\overline{\text{BL}}$), we get

$$J^T J Q_k = J^T U_{k+1} B_k = Q_k B_k^T B_k + \gamma_{k+1} q_{k+1} e_{k+1}^T B_k = Q_k T_k + \gamma_{k+1} \delta_{k+1} q_{k+1} e_k^T,$$

where

$$T_k = B_k^T B_k = \begin{bmatrix} \gamma_1^2 + \delta_2^2, & \gamma_2 \delta_2, & \dots, & 0, & 0 \\ \gamma_2 \delta_2, & \gamma_2^2 + \delta_3^2, & \dots, & 0, & 0 \\ -- & -- & -- & -- & -- \\ 0, & 0, & \dots, & \gamma_{k-1}^2 + \delta_k^2, & \gamma_k \delta_k \\ 0, & 0, & \dots, & \gamma_k \delta_k, & \gamma_k^2 + \delta_{k+1}^2 \end{bmatrix}$$

is a symmetric tridiagonal matrix of order k. Therefore, ($\overline{\text{SL}}$) holds, where $B = J^T J$, $T_k = B_k^T B_k$ and $\alpha_i = \gamma_i^2 + \delta_{i+1}^2$, $\beta_i = \gamma_i \delta_i$, $1 \leq i \leq k$ and we can use Theorem 41, which implies that the vectors q_i, $1 \leq i \leq k+1$ form a basis in the Krylov subspace $\mathcal{K}_i = \text{span}\ (g, Bg, \dots, B^{k-1}g)$, where $B = J^T J$ a $g = J^T f$. Using the first relation from ($\overline{\text{BL}}$), we obtain

$$U_{k+1}^T J Q_k = U_{k+1}^T U_{k+1} B_k$$

and the second relation from ($\overline{\text{BL}}$) gives

$$U_{k+1}^T J Q_k = B_k Q_k^T Q_k + \gamma_{k+1} e_{k+1} q_{k+1}^T Q_k = B_k,$$

so that $U_{k+1}^T U_{k+1} = I$ (The vectors u_i, $1 \leq i \leq k+1$, are mutually orthogonal with unit norms).

Remark 54 From proof of Theorem 43 it follows that the symmetric Lanzcos process defined by the SPD matrix $B \in R^{n \times n}$ and the vector $g \in R^n$ is equivalent to the Lanzcos bidiagonalization process defined by the matrix $J \in R^{m \times n}$ and the vector $f \in R^m$, if $B = J^T J$ and $g = J^T f$. The equivalence consists in the fact that both processes generate the same vectors q_i, $1 \leq i \leq k$, and $\alpha_i = \gamma_i^2 + \delta_{i+1}^2$, $\beta_i = \gamma_i \delta_i$, $1 \leq i \leq k$ hold, where k is an index such that $\alpha_i \neq 0$, $\beta_i \neq 0$, $\gamma_i \neq 0$, $\delta_i \neq 0$, $1 \leq i \leq k$.

Remark 55 The Lanczos bidiagonalization process can be used for solving the linear system $J^T J s + J^T g = 0$. We set $s_1 = 0$ and look for a vectors s_{i+1}, $1 \leq i \leq k$, that satisfy the relation

$$s_{i+1} = \arg \min_{s \in \mathcal{K}_i} \parallel Js + f \parallel .$$

Since $s \in \mathcal{K}_i$ if and only if $s = Q_i z$, where $z \in R^i$, we can write $s_{i+1} = Q_i z_i$, where

$$z_i = \arg \min_{z \in R^i} \parallel B_i z + \delta_1 e_1 \parallel$$

(it follows from the relations $f = U_{i+1}(\delta_1 e_1)$, $JQ_i = U_{i+1} B_i$ and $U_{i+1}^T U_{i+1} = I$). If $\gamma_{k+1} \delta_{k+1} = 0$, then the vector $s_{k+1} \in \mathcal{K}_k$ is a solution of the linear system $J^T J s + J^T f = 0$ (it follows from Remark 52 and Remark 54).

Remark 56 The vectors s_{i+1}, $1 \leq i \leq k$, defined in Remark 55 are identical with the vectors s_{i+1}, $1 \leq i \leq k$, generated by the CGNE method defined by the matrix $J \in R^{m \times n}$ and the vector $f \in R^m$ (it follows from Theorem 42 and Remark 54).

An advantage of the Lanczos bidiagonalization process is the fact, that the vectors s_{i+1}, $1 \leq i \leq k$, can be determined using stable operations (Givens elementary rotations). This is a basis for the LSQR method. The principle of the LSQR method consists in the sucessively computed decompositions

$$P_i B_i = \begin{bmatrix} R_i \\ 0 \end{bmatrix}, \qquad P_i(\delta_1 e_1) = \begin{bmatrix} h_i \\ \overline{\eta}_{i+1} \end{bmatrix},$$

where

$$R_i = \begin{bmatrix} \rho_1, & \sigma_2, & \ldots, & 0 \\ 0, & \rho_2, & \ldots, & 0 \\ 0, & 0, & \ldots, & \rho_i \end{bmatrix}, \qquad h_i = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \ldots \\ \eta_i \end{bmatrix}.$$

Here $P_i \in R^{i \times i}$, $1 \le i \le k$, are orthogonal matrices (products of Givens elementary rotations) and $R_i \in R^{i \times i}$, $1 \le i \le k$, are upper bidiagonal matrices. First we demonstrate two steps of this process. At the start of the first step we have the matrices

$$B_1 = \begin{bmatrix} \overline{\rho}_1 \\ \delta_2 \end{bmatrix}, \qquad \delta_1 e_1 = \begin{bmatrix} \overline{\eta}_1 \\ 0 \end{bmatrix},$$

where $\overline{\rho}_1 = \gamma_1$ and $\overline{\eta}_1 = \delta_1$. If we set

$$P_1 = \frac{1}{\sqrt{\overline{\rho}_1^2 + \delta_2^2}} \begin{bmatrix} \overline{\rho}_1, & \delta_2 \\ -\delta_2, & \overline{\rho}_1 \end{bmatrix},$$

we obtain

$$P_1 B_1 = \frac{1}{\sqrt{\overline{\rho}_1^2 + \delta_2^2}} \begin{bmatrix} \overline{\rho}_1^2 + \delta_2^2 \\ 0 \end{bmatrix} \triangleq \begin{bmatrix} \rho_1 \\ 0 \end{bmatrix},$$

$$P_1(\delta_1 e_1) = \frac{1}{\sqrt{\overline{\rho}_1^2 + \delta_2^2}} \begin{bmatrix} \overline{\rho}_1 \overline{\eta}_1 \\ -\delta_2 \overline{\eta}_1 \end{bmatrix} \triangleq \begin{bmatrix} \eta_1 \\ \overline{\eta}_2 \end{bmatrix}$$

and

$$P_1 \begin{bmatrix} 0 \\ \gamma_2 \end{bmatrix} = \frac{1}{\sqrt{\overline{\rho}_1^2 + \delta_2^2}} \begin{bmatrix} \delta_2 \gamma_2 \\ \overline{\rho}_1 \gamma_2 \end{bmatrix} \triangleq \begin{bmatrix} \sigma_2 \\ \overline{\rho}_2 \end{bmatrix}.$$

At the start of the second step we have the matrices

$$\begin{bmatrix} P_1, & 0 \\ 0, & 1 \end{bmatrix} \begin{bmatrix} \overline{\rho}_1, & 0 \\ \delta_2, & \gamma_2 \\ 0, & \delta_3 \end{bmatrix} = \begin{bmatrix} \rho_1, & \sigma_2 \\ 0, & \overline{\rho}_2 \\ 0, & \delta_3 \end{bmatrix}, \qquad \begin{bmatrix} P_1, & 0 \\ 0, & 1 \end{bmatrix} \begin{bmatrix} \overline{\eta}_1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \eta_1 \\ \overline{\eta}_2 \\ 0 \end{bmatrix}$$

and we can set

$$P_2 = \begin{bmatrix} 1, & 0 \\ 0, & \overline{P}_2 \end{bmatrix} \begin{bmatrix} P_1, & 0 \\ 0, & 1 \end{bmatrix}, \qquad \overline{P}_2 = \frac{1}{\sqrt{\overline{\rho}_2^2 + \delta_3^2}} \begin{bmatrix} \overline{\rho}_2, & \delta_3 \\ -\delta_3, & \overline{\rho}_2 \end{bmatrix}.$$

Continuing in this way, we get the recurrence relations

$$\overline{\rho}_1 = \gamma_1, \qquad \overline{\eta}_1 = \delta_1$$

and

$$\begin{aligned} \rho_i = \sqrt{\overline{\rho}_i^2 + \delta_{i+1}^2} \quad &, \quad \lambda_i = \frac{\overline{\rho}_i}{\rho_i}, \quad \tau_i = \frac{\delta_{i+1}}{\rho_i}, \\ \overline{\rho}_{i+1} = \lambda_i \gamma_{i+1} \quad &, \quad \sigma_{i+1} = \tau_i \gamma_{i+1}, \\ \eta_i = \lambda_i \overline{\eta}_i \quad &, \quad \overline{\eta}_{i+1} = -\tau_i \overline{\eta}_i, \end{aligned}$$

for $1 \le i \le k$. Now we derive the recurrence relations for the vectors s_{i+1}, $1 \le i \le k$. Since

$$P_i(B_i z + \delta_1 e_1) = \begin{bmatrix} R_i \\ 0 \end{bmatrix} z + \begin{bmatrix} h_i \\ \overline{\eta}_{i+1} \end{bmatrix}$$

and $P_i^T P_i = I$, we can set $s_{i+1} = Q_i z_i$, where

$$z_i = \arg\min_{z \in R^i} \| R_i z + h_i \| .$$

Since the matrix $R_i \in R^{i \times i}$ is nonsingular, $R_i z_i + h_i = 0$ has to hold. In spite of a simple structure of matrices R_i, $1 \le i \le k$, the vectors z_i, $1 \le i \le k$, and s_{i+1}, $1 \le i \le k$, can be determined recursively. First we demonstrate two steps of this process. At the start of the first step it holds

$$R_1 = [\rho_1], \qquad h_1 = [\eta_1]$$

and the vector $z_1 = [\zeta_{11}]^T$ can be determined from the formula

$$R_1 z_1 + h_1 = [\rho_1][\zeta_{11}] + [\eta_1] = 0,$$

which gives $\zeta_{11} = -\eta_1/\rho_1$. Therefore,

$$s_2 = \zeta_{11} q_1 = s_1 + \frac{\eta_1}{\rho_1} p_1,$$

where

$$s_1 = 0, \qquad p_1 = -q_1.$$

At the start of the second step it holds

$$R_2 = \begin{bmatrix} \rho_1, & \sigma_2 \\ 0, & \rho_2 \end{bmatrix}, \qquad h_2 = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix}$$

and the vector $z_2 = [\zeta_{21}, \zeta_{22}]$ can be determined from the formula

$$R_2 z_2 + h_2 = \begin{bmatrix} \rho_1, & \sigma_2 \\ 0, & \rho_2 \end{bmatrix} \begin{bmatrix} \zeta_{21} \\ \zeta_{22} \end{bmatrix} + \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = 0,$$

which gives $\zeta_{22} = -\eta_2/\rho_2$ and $\zeta_{21} = -\eta_1/\rho_1 + \zeta_{22}\sigma_2/\rho_1$. Therefore,

$$s_3 = \zeta_{21} q_1 + \zeta_{22} q_2 = \zeta_{21} q_1 + \zeta_{22} \left(q_2 - \frac{\sigma_2}{\rho_1} q_1 \right) = s_2 + \frac{\eta_2}{\rho_2} p_2,$$

where

$$p_2 = -q_2 + \frac{\sigma_2}{\rho_1} p_1.$$

Continuing in this way, we obtain the recurrence relations

$$s_1 = 0, \qquad p_1 = -q_1$$

and

$$\begin{aligned} s_{i+1} &= s_i + \frac{\eta_i}{\rho_i} p_i, \\ p_{i+1} &= -q_{i+1} + \frac{\sigma_{i+1}}{\rho_i} p_i, \end{aligned}$$

for $1 \le i \le k$.

Definition 30 Let $J \in R^{m \times n}$ be a matrix with linearly independent columns and $f \in R^m$. Then the iterative process described by the recurrence relations

$$s_1 = 0, \qquad \delta_1 u_1 = f, \qquad \gamma_1 q_1 = J^T u_1, \qquad p_1 = q_1$$

and

$$\delta_{i+1} u_{i+1} = J q_i - \gamma_i u_i,$$

$$\gamma_{i+1} q_{i+1} = J^T u_{i+1} - \delta_{i+1} q_i,$$

$$\rho_i = \sqrt{\overline{\rho}_i^2 + \delta_{i+1}^2}, \qquad \lambda_i = \frac{\overline{\rho}_i}{\rho_i}, \qquad \tau_i = \frac{\delta_{i+1}}{\rho_i},$$

$$\overline{\rho}_{i+1} = \lambda_i \gamma_{i+1}, \qquad \sigma_{i+1} = \tau_i \gamma_{i+1},$$

$$\eta_i = \lambda_i \overline{\eta}_i, \qquad \overline{\eta}_{i+1} = -\tau_i \overline{\eta}_i,$$

$$s_{i+1} = s_i + \frac{\eta_i}{\rho_i} p_i,$$

$$p_{i+1} = -q_{i+1} + \frac{\sigma_{i+1}}{\rho_i} p_i,$$

for $1 \le i \le n$, where the coefficients γ_i, δ_i, $1 \le i \le n$, are chosen for $u_i \in R^m$, $q_i \in R^n$, $1 \le i \le n$ to have unit norms, is the LSQR method defined by the matrix $J \in R^{m \times n}$ and the vector $f \in R^m$.

The LSQR method can be used for a realization of an inexact trust region method in the same way as the CGNE (or CG) method since, by Remark 56, both these methods generate the same vectors s_{i+1}, $1 \le i \le k$, where $k \le n$ and $J^T J s_{k+1} + J^T f = 0$. Now we show the possibility for estimating residual norms.

Theorem 44 Let $s_{i+1} \in R_n$, γ_{i+1}, δ_{i+1}, $\rho_i > 0$, η_i, $1 \le i \le k$, be quantities generated by the LSQR method. Then

$$\| J^T(J s_{i+1} + f) \| = \gamma_{i+1} \delta_{i+1} \frac{|\eta_i|}{\rho_i}$$

holds for $1 \le i \le k$.

Proof Let $\gamma_{i+1} \ne 0$, $\delta_{i+1} \ne 0$. Then using $(\overline{\mathrm{BL}})$ and Remark 55 we get

$$\begin{aligned} J^T(J s_{i+1} + f) &= J^T(J Q_i z_i + f) = J^T U_{i+1}(B_i z_i + \delta_1 e_1) = \\ &= (Q_i B_i^T + \gamma_{i+1} q_{i+1} e_{i+1}^T)(B_i z_i + \delta_1 e_1) = \gamma_{i+1} q_{i+1} e_{i+1}^T B_i z_i = \\ &= \gamma_{i+1} \delta_{i+1} q_{i+1} e_i^T z_i, \end{aligned}$$

since $B_i^T(B_i z_i + \delta_1 e_1) = 0$ by the definition of the vector z_i, $e_{i+1}^T e_1 = 0$ and $e_{i+1}^T B_i = \delta_{i+1} e_i^T$. But $Q_i^T Q_i = I$ and, therefore, $Q_i^T s_{i+1} = Q_i^T Q_i z_i = z_i$, so that $e_i^T z_i = e_i^T Q_i^T s_{i+1} = q_i^T s_{i+1}$, which together with $\| q_{i+1} \| = 1$ gives

$$\| J^T(J s_{i+1} + f) \| = \gamma_{i+1} \delta_{i+1} |q_i^T s_{i+1}|.$$

But

$$q_i^T s_{i+1} = q_i^T s_i + \frac{\eta_i}{\rho_i} q_i^T p_i = q_i^T Q_{i-1} z_{i-1} - \frac{\eta_i}{\rho_i} q_i^T q_i + \frac{\eta_i \sigma_i}{\rho_i \rho_{i-1}} q_i^T p_{i-1} = -\frac{\eta_i}{\rho_i},$$

since $q_i^T Q_{i-1} = 0$, $q_i^T q_i = 1$ and the vector p_{i-1} is a linear combination of columns of the matrix Q_{i-1}, and, therefore, $q_i^T p_{i-1} = 0$. If $\gamma_{i+1} = 0$, $\delta_{i+1} = 0$, then $\| J^T(J s_{i+1} + f) \| = 0$ holds (Remark 55).

Theorem 44 can be used for termination of the iterative process (it is not necessary to compute the residual vector $\| J^T(J s_{i+1} + f) \|$).

The following table shows a comparison of several realizations of the inexact QN method for nonlinear least squares (CG - conjugate gradient method with sparse Hessian approximation, CGNE - the CGNE method with sparse Jacobian representation, LSQR - the LSQR method with sparse Jacobian representation) by means of 22 test problems with 100 variables (the total number of iterations NIT, function evaluations NFV and gradient evaluations NFG, are given together with the number of failures and the total computational time).

Method	MT-NFV-NFG	fail	time
GN + CG	1584-1819-1603	–	31.47
GN + CGNE	1602-1835-1621	–	43.67
GN + LSQR	1358-1584-1377	–	51.08

The above table implies some conclusions. If the rows of the Jacobian matrix are not dense, then it is more advantageuos to use the sparse Hessian representation together with the CG method (GN+CG), since the number of nonzero elements of the matrix B is usualy much lower then the number of nonzero elements of the matrix J in this case. Otherwise we make a decision from the complexity of the objective function. The CGNE method uses simpler matrix operations but the LSQR method requires a lower number of iterations and function evaluations.

5 Methods for solving systems of nonlinear equations

Let $f : R^n \to R^n$ be a continuously differentiable mapping. We want to find a point $x^* \in R^n$, such that $f(x^*) = 0$. A great number methods based on various ideas were developed for solving this problem. Here we restrict ourselves to methods closely related to the optimization methods, which are usually simple and efficient. We do not consider methods based on homotopies, simplicial methods and methods which require solution of systems of differential equations.

Similarity between methods for solving systems of nonlinear equations and optimization methods follows from the following facts:

- Optimization methods can be considered as methods for solution of the system $g(x) = 0$, where $g : R^n \to R^n$ is the gradient of the objective function $F : R^n \to R$. Since the Jacobian matrix of the mapping g is the Hessian matrix of the function F, i.e. it is symmetric (under standard conditions on the function F), we deal in this case with rather special nonlinear systems.
- Solution of the nonlinear system can be transformed to the minimization of the function $F : R^n \to R$ defined by the relation $F(x) = (1/2) \parallel f(x) \parallel^2$ (sum of squares). However, we can obtain a local minimum of the function $F : R^n \to R$, which is not a solution to the system $f(x) = 0$.

The following theorem gives a relation between the local minima of the function $F(x) = (1/2) \parallel f(x) \parallel^2$ and the solutions to the system $f(x) = 0$.

Theorem 45 Let $f : R^n \to R^n \in C^1$ and let the point $x^* \in R^n$ be a local minimizer of the function $F(x) = (1/2) \parallel f(x) \parallel^2$, such that the Jacobian matrix $J(x^*)$ of the mapping $f : R^n \to R^n$ is nonsingular. Then x^* is a global minimizer of the function F and $f(x^*) = 0$.

Proof The gradient of the function F at the point x^* can be expressed in the form

$$g(x^*) = J^T(x^*)f(x^*).$$

Since the matrix $J(x^*)$ is nonsingular, we can write

$$f(x^*) = (J^T(x^*))^{-1}g(x^*),$$

so that $f(x^*) = 0$ if and only if $g(x^*) = 0$. But $g(x^*) = 0$ implies stationarity of the function F. Since $F(x) \geq 0 = F(x^*)$, the point x^* is a global minimizer of the function F and $f(x^*) = 0$.

In the sequel, we will use the following conditions for the Jacobian matrix of the mapping $f : R^n \to R^n$.

1. Uniform boundedness:

$$\parallel J(x) \parallel \leq \overline{J} \tag{J3}$$

$\forall x \in R^n$.

2. Uniform regularity:

$$\| J^{-1}(x) \| \leq 1/\underline{J} \tag{J4}$$

$\forall x \in R^n$.

3. Lipschitz continuity:

$$\| J(x+d) - J(x) \| \leq \overline{L} \| d \| \tag{J5}$$

$\forall x \in R^n$, $\forall d \in R^n$.

These conditions have similar meaning as the conditions (F3)-(F5) for the objective function $F : R^n \rightarrow R$. If the condition (J3) holds, then the condition (J4) can be replaced by the following equivalent condition

$$\kappa(J(x)) \leq \overline{J}/\underline{J} \tag{$\overline{\mathrm{J4}}$}$$

$\forall x \in R^n$, where $\kappa(J(x))$ is a spectral condition number of the matrix $J(x)$.

Similarly to a basic optimization method (Section 1.4), we can define a basic method for solving systems of nonlinear equations, which is an iterative process generating the sequence $x_i \in R^n$, $i \in N$, such that

$$x_{i+1} = x_i + \alpha_i s_i$$

where the direction vector $s_i \in R^n$ is determined from the values x_j, f_j, J_j, $1 \leq j \leq i$, and the stepsize $\alpha_i > 0$ is determined from the behaviour of the function $F(x) = (1/2) \| f(x) \|^2$ in the neighborhood of the point $x_i \in R^n$.

Definition 31 We say that the basic method for solving systems of nonlinear equations is globally convergent, if

$$\lim_{i \rightarrow \infty} \| f(x_i) \| = 0$$

holds for an arbitrary initial point $x_1 \in R^n$.

The best known method for solving systems of nonlinear equations is the Newton method. This method is defined by the formulas

$$\begin{aligned} s_i &= -J^{-1}(x_i) f(x_i), \\ \alpha_i &= 1. \end{aligned}$$

The direction vector determined by the Newton method is identical with the direction vector obtained by the Gauss-Newton method applied to the sum of squares $F(x) = (1/2) \| f(x) \|^2$, since (if (J4) is satisfied) it holds

$$(J^T(x_i) J(x_i))^{-1} J^T(x_i) = J^{-1}(x_i).$$

The matrix $B_i = J^T(x_i) J(x_i)$ is positive definite in this case, so that the Newton method for solving systems of nonlinear equations can be realized as a line search method (in contrast to the Newton method for unconstrained optimization described in Section 3.7).

In the subsequent text, we will consider methods, which use approximations A_i of the Jacobian matrices $J_i = J(x_i)$, $i \in N$, such that

$$\|A_i - J_i\| \leq \overline{\vartheta}, \tag{A3}$$

$$\|A_i^{-1}\| \leq 1/\underline{A} \tag{A4}$$

hold, where $\overline{\vartheta} > 0$ and $\underline{A} > 0$.

Lemma 14 Let the assumptions (J3)-(J4) be satisfied and let $\gamma > 0$. Then numbers $\overline{\vartheta} > 0$, $\underline{A} > 0$, $\overline{\vartheta} \le \gamma \underline{A}$ exist, such that matrices A_i, $i \in N$, satisfying conditions (A3)-(A4), can be found.

Proof Let $\underline{J}$ and $\overline{J}$ be constants from (J3)-(J4). Denote $\kappa = \overline{J}/\underline{J}$, $\lambda = 2\kappa\gamma/(1+2\kappa\gamma)$ and set $2\overline{\vartheta} = \underline{J}\lambda/\kappa$. Let A_i be a matrix satisfying the condition (A3). Since $A_i s = J_i s + (A_i - J_i)s$, we can write

$$\begin{aligned} \|A_i s\|^2 &= s^T J_i^T J_i s + 2s^T (A_i - J_i)^T J_i s + \|(A_i - J_i)s\|^2 \\ &\ge \underline{J}^2 \|s\|^2 - 2\overline{\vartheta}\overline{J}\|s\|^2 = (1-\lambda)\underline{J}^2\|s\|^2 \end{aligned}$$

$\forall s \in R^n$, so that $\|A_i^{-1}\| \le 1/\underline{A}$, where $\underline{A} = \underline{J}\sqrt{1-\lambda} > 0$ (since $0 < \lambda < 1$). Moreover, $\lambda = 2\kappa\gamma(1-\lambda)$, so that

$$0 < 2\overline{\vartheta} = \underline{J}\lambda/\kappa = 2\underline{J}\gamma(1-\lambda) < 2\underline{J}\gamma\sqrt{1-\lambda} = 2\gamma\underline{A},$$

which was necessary to prove.

5.1 Line search methods

For description of line search methods for solving systems of nonlinear equations, we use the nontation $h_i = A_i^T f_i$ for approximation of the gradient $g_i = J_i^T f_i$.

Definition 32 We say that the basic method for solving systems of nonlinear equations $x_{i+1} = x_i + \alpha_i s_i$, $i \in N$ is a line search method if:
1) The direction vectors $s_i \in R^n$, $i \in N$, are determined in such a way that

$$\|A_i s_i + f_i\| \le \overline{\omega}\|f_i\|, \qquad (\overline{\mathrm{S1}})$$

where $0 \le \overline{\omega} < 1$.
2) The stepsizes $\alpha_i > 0$, $i \in N$, are selected, so that α_i is the first member of the sequence α_i^j, $j \in N$ (where $\alpha_i^1 = 1$ and $\underline{\beta}\alpha_i^j \le \alpha_i^{j+1} < \alpha_i^j$ $\forall j \in N$), such that

$$F_{i+1} - F_i \le \underline{\rho}\alpha_i h_i^T s_i, \qquad (\overline{\mathrm{S2a}})$$

where $0 < \underline{\beta} < 1$ and $0 < \underline{\rho} < 1$ (notice that ($\overline{\mathrm{S1}}$) implies $h_i^T s_i = f_i^T A_i s_i < 0$).

Lemma 15 Let the function $f : R^n \to R^n$ satisfy the assumptions (J3)-(J5). Let the matrices A_i, $i \in N$, satisfy (A3)-(A4) with $\overline{\vartheta} \le \gamma\underline{A}$, where $\gamma = (1-\overline{\omega})/(1+\overline{\omega})$ (satisfying these conditions guarantees Lemma 14) and let $\underline{\rho} < (\gamma\underline{A} - \overline{\vartheta})/(\gamma\underline{A})$. Then a direction vector $s_i \in R^n$ satisfying the condition ($\overline{\mathrm{S1}}$) and stepsize $\alpha_i > 0$ satisfying the condition ($\overline{\mathrm{S2a}}$) can be found in every iteration. Moreover a constant $\underline{\alpha} > 0$ exists, such that $\alpha_i \ge \underline{\alpha}$ $\forall i \in N$.

Proof Existence of a direction vector $s_i \in R^n$ satisfying the condition ($\overline{\mathrm{S1}}$) follows immediately from (A4) (if the matrix A_i is nonsingular, we can choose a vector s_i in such a way that $\|A_i s_i + f_i\| = 0$). Using ($\overline{\mathrm{S1}}$), we obtain

$$(1-\overline{\omega})\|f_i\| \le \|A_i s_i\| \le (1+\overline{\omega})\|f_i\|,$$

which together with (A4) gives

$$\|s_i\| \le \frac{1}{\underline{A}}\|A_i s_i\| \le \frac{1+\overline{\omega}}{\underline{A}}\|f_i\| = \frac{1-\overline{\omega}}{\gamma\underline{A}}\|f_i\|,$$

so that

$$-h_i^T s_i = -f_i^T(A_i s_i + f_i) + f_i^T f_i \ge (1-\overline{\omega})\|f_i\|^2 \ge \gamma\underline{A}\|f_i\|\|s_i\|. \qquad (*)$$

Since $g_i^T s_i = h_i^T s_i - f_i^T(A_i - J_i)s_i$, we can write

$$g_i^T s_i \le h_i^T s_i + \overline{\vartheta}\|f_i\|\|s_i\|,$$

which together with the previous inequality gives

$$g_i^T s_i \le \frac{\gamma \underline{A} - \overline{\vartheta}}{\gamma \underline{A}} h_i^T s_i < 0.$$

From the inequality $g_i^T s_i < 0$, the existence of a stepsize $\underline{\alpha} > 0$ satisfying the condition $F_{i+1} - F_i \le \varepsilon_1 \alpha_i g_i^T s_i$ for an arbitrary constant $0 < \varepsilon_1 < 1$ follows (Lemma 1). Using the inequality between $g_i^T s_i$ and $h_i^T s_i$ and setting $\underline{\rho} = \varepsilon_1(\gamma \underline{A} - \overline{\vartheta})/(\gamma \underline{A})$, we obtain ($\overline{\text{S2a}}$). Then either $\alpha_i = \alpha_i^1 = 1$ or $\alpha_i = \alpha_i^k = \beta \alpha_i^{k-1}$ holds, where $\underline{\beta} \le \beta < 1$ and $F(x_i + \alpha_i^{k-1} s_i) - F(x_i) \ge \underline{\rho} \alpha_i^{k-1} h_i^T s_i$. If $\alpha_i < 1$, then we can write

$$F(x_i + \frac{\alpha_i}{\beta} s_i) - F(x_i) \ge \underline{\rho} \frac{\alpha_i}{\beta} h_i^T s_i.$$

On the other hand, using the main value theorem together with the assumptions (J3)-(J5), we obtain

$$\begin{aligned} F(x_i + \frac{\alpha_i}{\beta} s_i) - F(x_i) &= \frac{\alpha_i}{\beta} s_i^T g(x_i + \mu \frac{\alpha_i}{\beta} s_i) \\ &\le \frac{\alpha_i}{\beta} \left(g_i^T s_i + \|s_i\| \|g(x_i + \mu \frac{\alpha_i}{\beta} s_i) - g(x_i)\| \right) \\ &\le \frac{\alpha_i}{\beta} \left(\frac{\gamma \underline{A} - \overline{\vartheta}}{\gamma \underline{A}} h_i^T s_i + \frac{\alpha_i}{\beta} (\overline{J}^2 + \overline{LF}) \|s_i\|^2 \right), \end{aligned}$$

since $0 \le \mu \le 1$ and for $d_i = (\alpha_i/\beta) s_i$ it holds

$$\begin{aligned} \|g(x_i + \mu d_i) - g(x_i)\| &= \|J^T(x_i + \mu d_i) f(x_i + \mu d_i) - J^T(x_i) f(x_i)\| \\ &\le \|J^T(x_i + \mu d_i)(f(x_i + \mu d_i) - f(x_i))\| \\ &\quad + \|(J^T(x_i + \mu d_i) - J^T(x_i)) f(x_i)\| \\ &\le \overline{J} \|(f(x_i + \mu d_i) - f(x_i)\| + \overline{L} \mu \|d_i\| \|f_i\| \\ &= \overline{J} \| \int_0^1 J(x_i + \tau \mu d_i) \mu d_i d\tau \| + \overline{L} \mu \|d_i\| \|f_i\| \\ &\le (\overline{J}^2 + \overline{LF}) \|d_i\| \end{aligned}$$

($\overline{F}$ is an arbitrary constant, such that $\overline{F} \ge \|f_1\|$). Coupling both these equations and using (*), we obtain

$$\frac{\alpha_i}{\beta} (\overline{J}^2 + \overline{LF}) \|s_i\|^2 \ge (\gamma \underline{A} - \overline{\vartheta} - \underline{\rho} \gamma \underline{A}) \|f_i\| \|s_i\|,$$

which together with $\beta \ge \underline{\beta}$ and with the inequality connecting the norms $\|s_i\|$ and $\|f_i\|$ gives $\alpha_i \ge \underline{\alpha}$, where

$$\underline{\alpha} = \min \left(1, \frac{\underline{\beta}(\gamma \underline{A} - \overline{\vartheta} - \underline{\rho} \gamma \underline{A}) \underline{A}}{(1 + \overline{\omega})(\overline{J}^2 + \overline{LF})} \right).$$

Remark 57 Lemma 15 show that a sufficiently accurate solution to the linear system $A_i s_i + f_i = 0$ does not suffice, in contrast of the unconstrained optimization, but also that the Jacobian matrix J_i has to be sufficiently accurately approximated (the inequality $\overline{\vartheta} \le \gamma \underline{A}$). This is caused by the fact that, in the relation for the gradient of the function $F : R^n \to R$ at the point $x_i \in R^n$, the Jacobian matrix J_i is used, which differs from the matrix A_i (inexact approximation of the Jacobian matrix implies inaccurate gradient).

Theorem 46 (global convergence). Let the assumption of Lemma 15 be satisfied. Let $x_i \in R^n$, $i \in N$ be a sequence generated by the line search method ($\overline{\text{S1}}$)-($\overline{\text{S2}}$). Then $x_i \to x^\star$ and $f(x^\star) = 0$.

Proof Using (A3), we obtain

$$\|A_i s_i\| \le \|J_i s_i\| + \|(A_i - J_i)s_i\| \le (\overline{J} + \overline{\vartheta})\|s_i\|,$$

which together with ($\overline{\text{S2a}}$) and with the inequalities obtained in the proof of Lemma 15 gives

$$\begin{aligned} \|f_i\|(\|f_{i+1}\| - \|f_i\|) &\le \frac{1}{2}(\|f_{i+1}\| + \|f_i\|)(\|f_{i+1}\| - \|f_i\|) = F_{i+1} - F_i \\ &\le \underline{\rho}\alpha_i h_i^T s_i \le -\underline{\rho}\gamma\underline{A}\alpha\|f_i\|\|s_i\| \le -\underline{\rho}\gamma\underline{A}\alpha\frac{1}{\overline{J} + \overline{\vartheta}}\|f_i\|\|A_i s_i\| \\ &\le -\underline{\rho}\gamma\underline{A}\alpha\frac{1-\overline{\omega}}{\overline{J} + \overline{\vartheta}}\|f_i\|^2. \end{aligned}$$

Therefore, it holds

$$\|f_{i+1}\| \le \left(1 - \underline{\rho}\gamma\underline{A}\alpha\frac{1-\overline{\omega}}{\overline{J} + \overline{\vartheta}}\right)\|f_i\| \triangleq q\|f_i\|,$$

where $0 < q < 1$, so that

$$\sum_{i=1}^{\infty}\|f_i\| = \frac{1}{1-q}\|f_1\| < \infty,$$

which implies $f_i \to 0$. From the inequality connecting the norms $\|s_i\|$ and $\|f_i\|$ (see proof of Lemma 15) we obtain

$$\sum_{i=1}^{\infty}\|s_i\| \le \frac{1+\overline{\omega}}{\underline{A}}\sum_{i=1}^{\infty}\|f_i\| < \infty,$$

so that the sequence x_i, $i \in N$, satisfies the Bolzano-Cauchy condition. Therefore, $x_i \to x^\star$ which together with $f_i \to 0$ gives $f(x^\star) = 0$.

Remark 58 From the inequality $\| f_{i+1} \| \le q \| f_i \|$, $i \in N$, where $0 < q < 1$, it follows that $x_i \to x^\star$ R-superlinearly.

Remark 59 Statements of Lemma 15 and Theorem 46 remain valid if we replace the condition ($\overline{\text{S2a}}$) by one of the conditions

$$F_{i+1} - F_i \le -\underline{\rho}\alpha_i F_i \qquad (\overline{\text{S2b}})$$

or

$$\|f_{i+1}\| - \|f_i\| \le -\underline{\rho}\alpha_i\|f_i\| \qquad (\overline{\text{S2c}}),$$

where now $0 < \underline{\rho} < 2(1-\overline{\omega})(\gamma\underline{A} - \overline{\vartheta})/(\gamma\underline{A})$. It follows from the inequality

$$g_i^T s_i \le -(\gamma\underline{A} - \overline{\vartheta})\|f_i\|\|s_i\| \le -\frac{(1-\overline{\omega})(\gamma\underline{A} - \overline{\vartheta})}{\gamma\underline{A}}\|f_i\|^2 = -\frac{2(1-\overline{\omega})(\gamma\underline{A} - \overline{\vartheta})}{\gamma\underline{A}}F_i,$$

which can be used in the same way as ($\overline{\text{S2a}}$) in the proof of Lemma 15 and from the inequality $(F_{i+1} - F_i)/F_i = (\|f_{i+1}\|^2 - \|f_i\|^2)/\|f_i\|^2$, which gives

$$2(\|f_{i+1}\| - \|f_i\|)/\|f_i\| \le (F_{i+1} - F_i)/F_i \le (\|f_{i+1}\| - \|f_i\|)/\|f_i\|.$$

Theorem 47 (superlinear convergence). Let $x_i \in R^n$, $i \in N$, be a sequence obtained by a line search method, such that $x_i \to x^*$. Let the assumptions (J3)-(J5) be satisfied. Let $\alpha_i = 1$, whenever this value satisfies the condition ($\overline{\text{S2c}}$) (or ($\overline{\text{S2b}}$) or ($\overline{\text{S2a}}$)). Let

$$\lim_{i\to\infty}\frac{\| A_i s_i + f_i \|}{\| f_i \|} = 0 \qquad (\alpha)$$

and

$$\lim_{i\to\infty} \frac{\| (A_i - J_i)s_i \|}{\| s_i \|} = 0. \qquad (\beta)$$

Then an index $k \in N$ exists, such that $\alpha_i = 1$, $\forall i \geq k$ and the sequence x_i, $i \in N$, converges Q-superlinearly to the point $x^* \in R^n$.

Proof The proof will be slightly more general, since the obtained results will be used in the proof of Theorem 57. That means that in part (a)-(b) we suppose only that

$$\limsup_{i\to\infty} \frac{\| A_i s_i + f_i \|}{\| f_i \|} \leq \overline{\omega} < 1 \qquad (\overline{\alpha}).$$

If (α) holds, then we can set $\overline{\omega} = 0$ in all subsequent formulas
(a) We show that an index $k_1 \in N$ exists, such that

$$\| f_i \| (1 - \overline{\omega})/\overline{J} \leq \| s_i \| \leq \| f_i \| (1 + \overline{\omega})/\underline{J}$$

$\forall i \geq k_1$, if $\| J^* \| < \overline{J}$ and $\| (J^*)^{-1} \| < 1/\underline{J}$. Denote $\omega_i = (A_i s_i + f_i)/ \| f_i \|$ and $\vartheta_i = (A_i - J_i)s_i / \| s_i \|$. Then

$$J_i s_i = (A_i s_i + f_i) - (A_i - J_i)s_i - f_i = \omega_i \| f_i \| - \vartheta_i \| s_i \| - f_i.$$

so that

$$\| s_i \| \geq \frac{1 - \| \omega_i \|}{\| J_i \| + \| \vartheta_i \|} \| f_i \|$$

and since $\| \omega_i \| \leq \overline{\omega}$ and $\| \vartheta_i \| \to 0$ (by $(\overline{\alpha})$ and (β)) and $\| J_i \| \to \| J^* \| < \overline{J}$, an index $k_0 \in N$ exists, such that $\| s_i \| \geq \| f_i \| (1 - \overline{\omega})/\overline{J}$ $\forall i \geq k_0$. Similarly we obtain

$$s_i = J_i^{-1}(\omega_i \| f_i \| - \vartheta_i \| s_i \| - f_i),$$

so that

$$\| s_i \| \leq \frac{\| J_i^{-1} \| (1 + \| \omega_i \|)}{1 - \| J_i^{-1} \| \| \| \vartheta_i \|} \| f_i \|$$

and since $\| \omega_i \| \leq \overline{\omega}$ and $\| \vartheta_i \| \to 0$ (podle $(\overline{\alpha})$ by (β)) and $\| J_i^{-1} \| \to \| (J^*)^{-1} \| < 1/\underline{J}$, an index $k_1 \geq k_0$ exists, such that $\| s_i \| \leq \| f_i \| (1 + \overline{\omega})/\underline{J}$ $\forall i \geq k_1$.
(b) We show that an index $k \geq k_1$ exists, such that the value $\alpha_i = 1$ satisfies condition $(\overline{\text{S2c}})$, if $\underline{\rho} < 1 - \overline{\omega}$ (we proceed analogously in the case of condition $(\overline{\text{S2b}})$ or condition $(\overline{\text{S2a}})$). Using the mean value, we obtain

$$f(x_i + s_i) = f_i + J_i s_i + o(\| s_i \|) = (A_i s_i + f_i) - (A_i - J_i)s_i + o(\| s_i \|)$$

or

$$\frac{\| f(x_i + s_i) \|}{\| f_i \|} \leq \| \omega_i \| + \| \vartheta_i \| (1 + \overline{\omega})/\underline{J} + o(\| f_i \|)/ \| f_i \|,$$

so that $\limsup_{i\to\infty} \| f(x_i + s_i) \| / \| f_i \| \leq \overline{\omega}$ (by $(\overline{\alpha})$ and (β)). If $\underline{\rho} < 1 - \overline{\omega}$, then an index $k \geq k_1$ exists, such that condition $(\overline{\text{S2c}})$ with $\alpha_i = 1$ holds $\forall i \geq k$ (if (α holds), then the number $0 < \underline{\rho} < 1/2$ can be arbitrary, since $\| f(x_i + s_i) \| / \| f_i \| \to 0$).
(c) Suppose now that (α) holds. Using the mean value theorem, we obtain

$$\frac{\| x_{i+1} - x^* \|}{\| x_i - x^* \|} \leq \frac{\overline{J}}{\underline{J}} \frac{\| f_{i+1} \|}{\| f_i \|},$$

so that (α), (β) and (c) imply

$$\lim_{i\to\infty} \frac{\| x_{i+1} - x^* \|}{\| x_i - x^* \|} = \lim_{i\to\infty} \frac{\overline{J}}{\underline{J}} (\| \omega_i \| + \| \vartheta_i \| (1 + \overline{\omega})/\underline{J} + o(\| f_i \|)/ \| f_i \|) = 0$$

and $x^* \to x$ Q-superlinearly.

5.2 Trust region methods

In description of trust region methods for solving systems of nonlinear equations, we use the notation $L_i(s) = \| A_i s + f_i \| - \| f_i \|$ for a function which locally approximates the difference $\| f(x_i+s) \| - \| f(x_i) \|$ and the notation $\rho_i(s) = (\| f(x_i+s) \| - \| f_i(x_i) \|)/L_i(s)$ for ratio of actual and predicted decrease of the norm $f : R^n \to R^n$.

Definition 33 We say that the basic method $x_{i+1} = x_i + \alpha_i s_i$, $i \in N$, for solving a system of nonlinear equations, is a trust region method if:
1) The direction vectors $s_i \in R^n$, $i \in N$, are determined in such a wat that

$$\| s_i \| \le \Delta_i \qquad , (\overline{\text{T1a}})$$

$$\| s_i \| < \Delta_i \Rightarrow \| A_i s_i + f_i \| \le \overline{\omega}_i \|f_i\|, \qquad (\overline{\text{T1b}})$$

$$-L_i(s_i) \ge 2\underline{\sigma} \| A_i s_i \|, \qquad (\overline{\text{T1c}})$$

where $0 \le \overline{\omega}_i \le \overline{\omega} < 1$ and $0 < \underline{\sigma} < 1/2$.
2) The stepsizes $\alpha_i \ge 0$, $i \in N$, are selected, so that

$$\rho_i(s_i) \le 0 \Rightarrow \alpha_i = 0, \qquad (\overline{\text{T2a}})$$

$$\rho_i(s_i) > 0 \Rightarrow \alpha_i = 1. \qquad (\overline{\text{T2b}})$$

3) The trust region radii $0 < \Delta_i \le \overline{\Delta}$, $i \in N$, are changed so that

$$\rho_i(s_i) < \underline{\rho} \Rightarrow \underline{\beta} \| s_i \| \le \Delta_{i+1} \le \overline{\beta} \| s_i \|, \qquad (\overline{\text{T3a}})$$

$$\rho_i(s_i) \ge \underline{\rho} \Rightarrow \Delta_i \le \Delta_{i+1} \le \overline{\Delta}, \qquad (\overline{\text{T3b}})$$

where $0 < \underline{\beta} < \overline{\beta} < 1$ and $0 < \underline{\rho} < 1/2$.

In the sequel, we use the notation N_1, N_2 and N_3 for the sets of indices, such that $\|s_i\| < \Delta_i$, $\rho_i(s_i) > 0$ and $\rho_i(s_i) \ge \underline{\rho}$, respectively.

Lemma 16 Let the mapping $f : R^n \to R^n$ satisfy the conditions (J3)-(J5). Let the matrices A_i, $i \in N$, satisfy the conditions (A3)-(A4) with $\overline{\vartheta} \le \gamma \underline{A}$, where $\gamma = (1-2\underline{\rho})\underline{\sigma}$ (satisfying these conditions guarantees Lemma 14). Let $x_i \in R^n$, $i \in N$, be a sequence of points generated by the trust region method ($\overline{\text{T1}}$)-($\overline{\text{T3}}$). Then a constant $\underline{c} > 0$ exists, such that

$$\|s_i\| \ge \underline{c}\|f_i\| \;\forall i \in N$$

Proof (a) Let $i \in N_1$. Then ($\overline{\text{T1b}}$) implies

$$\big|\|A_i s_i\| - \|f_i\|\big| \le \|A_i s_i + f_i\| \le \overline{\omega}\|f_i\|,$$

so that $(1-\overline{\omega})\|f_i\| \le \|A_i s_i\|$. On the other hand the condition (A3) gives

$$\|A_i s_i\| \le \|J_i s_i\| + \|(A_i - J_i)s_i\| \le (\overline{J} + \overline{\vartheta})\|s_i\|.$$

Coupling both these inequalities, we obtain

$$\|s_i\| \ge \frac{1-\overline{\omega}}{\overline{J}+\overline{\vartheta}}\|f_i\|.$$

(b) Let $i \notin N_1$ and $i \notin N_3$. From ($\overline{\text{T1c}}$) it follows that $L_i(s_i) \le 0$, so that

$$\begin{aligned} L_i(s_i)\|f_i\| &= (\|A_i s_i + f_i\| - \|f_i\|)\|f_i\| \ge \left(\|A_i s_i + f_i\|^2 - \|f_i\|^2\right) \\ &= 2\left(f_i^T A_i s_i + \frac{1}{2} s_i^T A_i^T A_i s_i\right) \triangleq 2Q_i(s_i). \end{aligned} \qquad (*)$$

If $\|f(x_i+s_i)\| \le \|f(x_i)\|$, then the inequality $\rho_i(s_i) < \underline{\rho}$ together with $(*)$ gives

$$\begin{aligned} F(x_i+s_i) - F(x_i) &= \frac{1}{2}\left(\|f(x_i+s_i)\|^2 - \|f(x_i)\|^2\right) \\ &\ge \left(\|f(x_i+s_i)\| - \|f(x_i)\|\right)\|f(x_i)\| \\ &\ge \underline{\rho}L_i(s_i)\|f_i\| \ge 2\underline{\rho}Q_i(s_i). \end{aligned}$$

If $\|f(x_i+s_i)\| \ge \|f(x_i)\|$, the above inequality holds trivially. Therefore, we can write

$$F(x_i+s_i) - F(x_i) \ge 2\underline{\rho}Q_i(s_i).$$

On the other hand, the assumptions (J3)-(J5) together with the mean value theorem give

$$\begin{aligned} \|g(x_i+\mu s_i) - g(x_i)\| &= \|J^T(x_i+\mu s_i)f(x_i+\mu s_i) - J^T(x_i)f(x_i)\| \\ &\le \|J^T(x_i+\mu s_i)(f(x_i+\mu s_i) - f(x_i))\| \\ &\quad +\|(J^T(x_i+\mu s_i) - J^T(x_i))f(x_i)\| \\ &\le \overline{J}\|(f(x_i+\mu s_i) - f(x_i)\| + \overline{L}\mu\|s_i\|\|f_i\| \\ &= \overline{J}\|\int_0^1 J(x_i+\tau\mu s_i)\mu s_i d\tau\| + \overline{L}\mu\|s_i\|\|f_i\| \\ &\le (\overline{J}^2 + \overline{LF})\|s_i\| \end{aligned}$$

for $0 \le \mu \le 1$, so that

$$\begin{aligned} F(x_i+s_i) - F(x_i) &\le g_i^T s_i + \|g(x_i+\mu s_i) - g(x_i)\|\|s_i\| \\ &\le g_i^T s_i + (\overline{J}^2 + \overline{LF})\|s_i\|^2 \\ &= f_i^T A_i s_i + f_i^T(J_i - A_i)s_i + (\overline{J}^2 + \overline{LF})\|s_i\|^2 \\ &\le Q_i(s_i) + \overline{\vartheta}\|s_i\|\|f_i\| + (\overline{J}^2 + \overline{LF})\|s_i\|^2. \end{aligned}$$

Coupling both these inequalities, we obtain

$$2\underline{\rho}Q_i(s_i) \le Q_i(s_i) + \overline{\vartheta}\|s_i\|\|f_i\| + (\overline{J}^2 + \overline{LF})\|s_i\|^2$$

or

$$-(1-2\underline{\rho})Q_i(s_i) \le \overline{\vartheta}\|s_i\|\|f_i\| + (\overline{J}^2 + \overline{LF})\|s_i\|^2.$$

The conditions ($\overline{\text{T1c}}$) and (A4) together with the inequality $(*)$ give

$$-Q_i(s_i) \ge -\frac{1}{2}L_i(s_i)\|f_i\| \ge \underline{\sigma}\|A_i s_i\|\|f_i\| \ge \underline{\sigma A}\|s_i\|\|f_i\|.$$

Substituting this relation into the previous inequality, we get

$$(1-2\underline{\rho})\underline{\sigma A}\|s_i\|\|f_i\| \le -(1-2\underline{\rho})Q_i(s_i) \le \overline{\vartheta}\|s_i\|\|f_i\| + (\overline{J}^2 + \overline{LF})\|s_i\|^2$$

or

$$\|s_i\| \ge \frac{(1-2\underline{\rho})\underline{\sigma A} - \overline{\vartheta}}{\overline{J}^2 + \overline{LF}}\|f_i\|$$

(the numerator is positive, since $\overline{\vartheta} < (1-2\underline{\rho})\underline{\sigma A}$).
(c) Let $i = 1$. If $\|f_1\| = 0$, then surely $\|s_1\| \ge \underline{c}\|f_1\|$ for an arbitrary constant $\underline{c} > 0$. If $\|f_1\| \ne 0$, then we obtain

$$\|s_1\| \ge \frac{\|s_1\|}{\|f_1\|}\|f_1\|.$$

(d) Let $i \notin N_1$, $i \in N_3$ and $i \neq 1$. Let $k < i$ be a maximum index, such that $k \notin N_1$, $k \in N_3$ and $k \neq 1$ do not hold simultaneously. Using ($\overline{\text{T3a}}$)-($\overline{\text{T3b}}$) and ($\overline{\text{T1a}}$), we can write

$$\|s_i\| = \Delta_i \geq \Delta_{k+1} \geq \min(\Delta_k, \underline{\beta}\|s_k\|) \geq \min\left(\|s_k\|, \underline{\beta}\|s_k\|\right) = \underline{\beta}\|s_k\|,$$

so that ($\overline{\text{T2a}}$)-($\overline{\text{T2b}}$) and (a)-(c) imply

$$\|s_i\| \geq \underline{\beta}\|s_k\| \geq \underline{c}\|f_k\| \geq \underline{c}\|f_i\|,$$

where

$$\underline{c} = \underline{\beta} \min\left(\frac{1-\overline{\omega}}{\overline{J}+\overline{\vartheta}}, \frac{(1-2\underline{\rho})\underline{\sigma A}-\overline{\vartheta}}{\overline{J}^2+\overline{LF}}, \frac{\|s_1\|}{\|f_1\|}\right).$$

Theorem 48 (global convergence). Let the assumptions of Lemma 16 be satisfied. Then $x_i \to x^\star$ and $f(x^\star) = 0$.

Proof (a) First we prove that $f_i \to 0$. Suppose the contrary. Since the sequence $\|f_i\|$, $i \in N$, is nonincreasing by ($\overline{\text{T2a}}$)-($\overline{\text{T2b}}$), a number $\underline{\varepsilon} > 0$ exists, such that $\|f_i\| \geq \underline{\varepsilon}$, $\forall i \in N$ and by Lemma 16 it holds

$$\|s_i\| \geq \underline{c\varepsilon}, \quad \forall i \in N.$$

Suppose first that the set N_3 is infinite. Since $N_3 \subset N_2$, we can write

$$\begin{aligned}\|f_i\| - \|f_{i+1}\| &= \|f(x_i)\| - \|f(x_i+s_i)\| \geq -\underline{\rho} L_i(s_i) \\ &\geq 2\underline{\rho\sigma}\|A_i s_i\| \geq 2\underline{\rho\sigma A c\varepsilon}, \; \forall i \in N_3.\end{aligned}$$

This implies

$$\begin{aligned}\|f_1\| &\geq \lim_{i\to\infty}(\|f_1\| - \|f_{i+1}\|) = \sum_{i=1}^{\infty}(\|f_i\| - \|f_{i+1}\|) \\ &\geq \sum_{i\in N_3}(\|f_i\| - \|f_{i+1}\|) \geq \sum_{i\in N_3} 2\underline{\rho\sigma A c\varepsilon} = \infty,\end{aligned}$$

which is the contradiction. Suppose now that the set N_3 is finite. Then ($\overline{\text{T3a}}$) implies $\Delta_i \to 0$, which together with ($\overline{\text{T1a}}$) gives $\|s_i\| \to 0$. But this is a contradiction to the inequality $\|s_i\| \geq \underline{c\varepsilon}$ $\forall i \in N$.
(b) Using ($\overline{\text{T1c}}$) we get $L_i(s_i) = \|A_i s_i + f_i\| - \|f_i\| \leq 0$, so that

$$\|f_i\| \geq \|A_i s_i + f_i\| \geq \|A_i s_i\| - \|f_i\|.$$

This inequality implies $\|A_i s_i\| \leq 2\|f_i\|$, so that

$$\underline{A}\|s_i\| \leq \|A_i s_i\| \leq 2\|f_i\|.$$

Now we show that $\sum_{i=1}^{\infty}\|s_i\| < \infty$. If the set N_3 is finite, then an index $l \notin N_3$ exists, such that $i \notin N_3$ $\forall i \geq l$. Therefore, it holds

$$\sum_{i=1}^{\infty}\|s_i\| \leq \sum_{i=1}^{l-1}\|s_i\| + \|s_l\|\sum_{i=l}^{\infty}\overline{\beta}^{i-l} \leq (l-1)\overline{\Delta} + \|s_l\|/(1-\overline{\beta}) < \infty$$

by ($\overline{\text{T3a}}$). If the set N_3 is infinite, then we can write (similarly as in (a))

$$\begin{aligned}\|f_1\| &\geq \sum_{i=1}^{\infty}(\|f_i\| - \|f_{i+1}\|) \geq \sum_{i\in N_3}(\|f_i\| - \|f_{i+1}\|) \\ &\geq 2\underline{\rho\sigma}\sum_{i\in N_3}\|A_i s_i\| \geq 2\underline{\rho\sigma A}\sum_{i\in N_3}\|s_i\|.\end{aligned}$$

Denote $N_3 = \{l_1, l_2, l_3, \ldots\}$. If we use Lemma 16, we obtain

$$\|s_{l_j+1}\| \leq \frac{2}{\underline{A}}\|f_{l_j+1}\| \leq \frac{2}{\underline{A}}\|f_{l_j}\| \leq \frac{2}{c\underline{A}}\|s_{l_j}\|$$

and ($\overline{\text{T3a}}$) implies $\|s_{l_j+k}\| \leq \overline{\beta}\|s_{l_j+k-1}\|$ $\forall 2 \leq k \leq l_{j+1} - l_j - 1$. Therefore, it holds

$$\begin{aligned}
\sum_{i=1}^{\infty}\|s_i\| &= \sum_{i=1}^{l_1-1}\|s_i\| + \sum_{j=1}^{\infty}\left[\|s_{l_j}\| + \sum_{k=1}^{l_{j+1}-l_j-1}\|s_{l_j+k}\|\right] \\
&\leq (l_1 - 1)\overline{\Delta} + \sum_{j=1}^{\infty}\|s_{l_j}\|\left[1 + \frac{2}{c\underline{A}}\sum_{k=1}^{l_{j+1}-l_j-1}\overline{\beta}^{k-1}\right] \\
&\leq (l_1 - 1)\overline{\Delta} + \left[1 + \frac{2}{c\underline{A}}\frac{1}{1-\overline{\beta}}\right]\sum_{i\in N_3}\|s_i\| \\
&\leq (l_1 - 1)\overline{\Delta} + \left[1 + \frac{2}{c\underline{A}}\frac{1}{1-\overline{\beta}}\right]\frac{\|f_1\|}{2\underline{\rho}\sigma\underline{A}} < \infty.
\end{aligned}$$

The inequality $\sum_{i=1}^{\infty}\|x_{i+1} - x_i\| \leq \sum_{i=1}^{\infty}\|s_i\| < \infty$ implies that the sequence x_i, $i \in N$, satisfies the Bolzano-Cauchy condition, so that $x_i \to x^\star$, which together with $f_i \to 0$ gives $f(x^\star) = 0$.

Theorem 49 (superlinear convergence). Let $x_i \in R^n$, $i \in N$, be a sequence generated by the trust region method ($\overline{\text{T1}}$) – ($\overline{\text{T3}}$), such that $x_i \to x^*$. Let the mapping $f : R^n \to R^n$ satisfy the conditions (J3)-(J5). Let

$$\lim_{i\to\infty}\overline{\omega}_i = 0 \tag{α}$$

and

$$\lim_{i\to\infty}\frac{\| (A_i - J_i)s_i \|}{\| s_i \|} = 0 \tag{β}$$

Then the sequence x_i, $i \in N$, converges Q-superlinearly to the point $x^* \in R^n$.

Proof (a) We show that an index $k_2 \in N$ exists, such that

$$-L_i(s_i) \geq 2\underline{\sigma}\underline{J} \| s_i \|$$

and

$$\| f_i \| \geq \frac{1}{2}\underline{J} \| s_i \|$$

$\forall i \geq k_2$, if $\underline{J} < 1/\| (J^*)^{-1} \|$. Denote $\vartheta_i = (A_i - J_i)s_i/\| s_i \|$. Then

$$\| A_i s_i \| = \| J_i s_i + \vartheta_i \| s_i \|\| \geq \| J_i s_i \| - \| \vartheta_i \|\| s_i \|$$

and since $\| \vartheta_i \| \to 0$, $J_i \to J^*$ and $\underline{J} < 1/\| (J^*)^{-1} \|$, an index $k_2 \in N$ exists, such that $\| A_i s_i \| \geq \underline{J} \| s_i \|$, $\forall i \geq k_2$. Using ($\overline{\text{T1c}}$), we can write

$$-L_i(s_i) \geq 2\underline{\sigma} \| A_i s_i \| \geq 2\underline{\sigma}\underline{J} \| s_i \| .$$

The definition of $L_i(s_i)$ and ($\overline{\text{T1c}}$) imply

$$0 \geq L_i(s_i) = \| A_i s_i + f_i \| - \| f_i \|$$

or

$$| \| A_i s_i \| - \| f_i \| | \leq \| A_i s_i + f_i \| \leq \| f_i \|,$$

so that $\| A_i s_i \| \leq 2 \| f_i \|$, which together with the inequality $\| A_i s_i \| \geq \underline{J} \| s_i \|$ gives $\| f_i \| \geq (\underline{J}/2) \| s_i \| \ \forall i \geq k_2$.
(b) We show that an index $k_3 \geq k_2$ exists, such that $i \in N_3 \ \forall i \geq k_3$. Using the mean value theorem we obtain

$$f(x_i + s_i) = f(x_i) + J_i s_i + o(\| s_i \|) = f(x_i) + A_i s_i - (A_i - J_i)\, s_i + o(\| s_i \|),$$

so that

$$\begin{aligned} \rho_i(s_i) &= \frac{\| f(x_i) \| - \| f(x_i + s_i) \|}{-L_i(s_i)} \geq \frac{-L_i(s_i) - \| \vartheta_i \| \| s_i \| + o(\| s_i \|)}{-L_i(s_i)} \geq \\ &\geq 1 - \frac{\| \vartheta_i \| \| s_i \| + o(\| s_i \|)}{2\underline{\sigma J} \| s_i \|} \to 1, \end{aligned}$$

since $\| \vartheta_i \| \to 0$. Since $\underline{\rho} < 1$, an index $k_3 \geq k_2$ exists, such that $\rho_i(s_i) \geq \underline{\rho} \ \forall i \geq k_3$.
(c) We show that an index $k \geq k_3$ exists, such that $i \in N_1 \ \forall i \geq k$. Notice first that the set $N_1 \subset N$ is infinite. Otherwise, $\| s_i \| \geq \Delta_i \geq \Delta_{k_3} \ \forall i \geq k_3$, has to hold, since (b) implies $i \in N_3 \ \forall i \geq k_3$. But this is a contradiction, since (a) implies $\| s_i \| \leq 2 \| f_i \| / \underline{J}$, so that $\| f_i \| \to 0$ implies $\| s_i \| \to 0$. Now we consider only indices $i \geq k_3$, $i \in N_1$ and denote $\omega_i = (A_i s_i + f_i)/\|s_i\|$. The conditions (α), (β) and $(\overline{\text{T1b}})$ imply $\| \omega_i \| \to 0$ and $\| \vartheta_i \| \to 0$, so that we can prove, by the same way as in proof of Theorem 47 (s $\overline{\omega} = 0$), that an index $k_4 \geq k_3$, $k_4 \in N_1$ exists, such that

$$\| f_i \| / \overline{J} \leq \| s_i \| \leq \| f_i \| / \underline{J}$$

$\forall i \geq k_4$, $i \in N_1$. Using the mean value theorem, we can write

$$f_{i+1} = f(x_i + s_i) = f_i + J_i s_i + o(\| s_i \|),$$

since $i \in N_3 \subset N_2$. Denote

$$\lambda_i = \frac{f_{i+1} - f_i - A_i s_i}{\| f_i \|}.$$

Then, using the previous considerations, we obtain $\| \lambda_i \| \leq \| \vartheta_i \| / \overline{J} + o(\| s_i \|) / \| s_i \| \to 0$. Since simultaneously $\| \omega_i \| \to 0$ holds, an index $k \geq k_4$, $k \in N_1$ exists, such that $\| \lambda_i \| < (\underline{J}/\overline{J})/2$ and $\| \omega_i \| < (\underline{J}/\overline{J})/2 \ \forall i \geq k$, $i \in N_1$. Then we can write

$$\begin{aligned} \| s_{i+1} \| &\leq \frac{1}{\underline{J}} \| f_{i+1} \| \leq \frac{1}{\underline{J}} (\| f_{i+1} - f_i - A_i s_i \| + \| A_i s_i + f_i \|) \leq \\ &\leq \frac{\overline{J}}{\underline{J}} (\| \lambda_i \| + \| \omega_i \|) \| s_i \| < \left(\frac{1}{2} + \frac{1}{2} \right) \| s_i \| = \| s_i \| . \end{aligned}$$

Since $i \in N_3$ by (b), then $\Delta_{i+1} \geq \Delta_i$ holds, which gives $\| s_{i+1} \| < \| s_i \| \leq \Delta_i \leq \Delta_{i+1}$, so that $i + 1 \in N_1$. Using induction, we obtain $i \in N_1 \ \forall i \geq k$.
(d) Superlinear convergence. It holds

$$\frac{\| f_{i+1} \|}{\| f_i \|} \leq \frac{\| f_{i+1} - f_i - A_i s_i \| + \| A_i s_i + g_i \|}{\| f_i \|} \leq \| \lambda_i \| + \| \omega_i \|$$

which together with $\| \lambda_i \| \to 0$ and $\| \omega_i \| \to 0$ gives

$$\lim_{i \to \infty} \frac{\| x_{i+1} - x^* \|}{\| x_i - x^* \|} \leq \frac{\overline{J}}{\underline{J}} \frac{\| f_{i+1} \|}{\| f_i \|} = 0.$$

5.3 Newton method

The Newton method utilizes the matrices $A_i = J(x_i)$, $\forall i \in N$, so that $\vartheta_i = (A_i - J_i)s_i / \| s_i \| = 0$, $\forall i \in N$.

Theorem 50 Let the assumptions (J3)-(J5) be satisfied and let $\| \omega_i \| = \| A_i s_i + f_i \| / \| f_i \| \le \overline{\omega} < 1$ $\forall i \in N$. Then the Newton method, realized either as a line search method or as a trust region method, is globally convergent. If $x_i \to x^*$ and $\| \omega_i \| \to 0$, then the rate of convergence is Q-superlinear.

Proof Global convergence follows immediately from Theorem 46 and Theorem 48. Superlinear convergence follows immediately from Theorem 47 and Theorem 49, since $\vartheta_i = 0$ $\forall i \in N$.

Remark 60 The Newton method for solving systems of nonlinear equations can be realized as a globally convergent line search method, which is not possible in the case of the Newton method for unconstrained optimization.

If the Jacobian matrices are not given analytically, we can utilize the discrete Newton method, which is based on difference approximation of the first order derivatives. But cancellation and round-off errors can violate the conditions (A3) and (A4) with $\overline{\vartheta} \le \gamma \underline{A}$ and γ determined by the selected realization. The following statements give an estimation of errors in the approximation of the Jacobian matrix.

Lemma 17 Let the assumption (J5) be satisfied and let

$$Ae_j = \frac{f(x + \delta e_j) - f(x)}{\delta} \tag{D}$$

hold for $1 \le j \le n$, where e_j, $1 \le j \le n$, are columns of the unit matrix of order n. Then

$$\| A - J(x) \| \le \frac{1}{2} \overline{L} \sqrt{n} \delta$$

Proof Using the mean value theorem, we obtain

$$f(x + \delta e_j) = f(x) + J(x)\delta e_j + \int_0^1 (J(x + \tau \delta e_j) - J(x))\delta e_j \, d\tau,$$

so that

$$\begin{aligned} \| (A - J(x))e_j \| &= \| \frac{f(x + \delta e_j) - f(x)}{\delta} - J(x)e_j \| \le \frac{1}{\delta} \| \int_0^1 (J(x + \tau \delta e_j) - J(x))\delta e_j \, d\tau \| \le \\ &\le \frac{1}{\delta} \frac{1}{2} \overline{L} \delta^2 \| e_j \|^2 = \frac{1}{2} \overline{L} \delta. \end{aligned}$$

Let $s \in R^n$ be an arbitrary vector with unit norm. Then

$$\begin{aligned} \| (A - J(x))s \| &= \| \sum_{j=1}^n (A - J(x))e_j e_j^T s \| \le \sum_{j=1}^n |e_j^T s| \, \| (A - J(x))e_j \| \le \frac{1}{2} \overline{L} \delta \sum_{j=1}^n |e_j^T s| \le \\ &\le \frac{1}{2} \overline{L} \sqrt{n} \delta \| s \| = \frac{1}{2} \overline{L} \sqrt{n} \delta \end{aligned}$$

and since

$$\| A - J(x) \| = \max_{\|s\|=1} \| (A - J(x))s \|,$$

the lemma is proved.

Theorem 51 Let the assumptions (J3)-(J5) be satisfied and let the matrix A be determined by the formula (D), where

$$\delta < \frac{\underline{J}\lambda}{\overline{L}\sqrt{n}\kappa}$$

and where κ, λ, γ are the numbers used in the proof of Lemma 14. Then $\|A - J(x)\| \le \overline{\vartheta}$, holds, where $\overline{\vartheta} \le \gamma \underline{A}$.

Proof From the Lemma 17, from the assumptions of the Theorem 51 and from the proof of the Lemma 14 it follows that

$$\|A - J(x)\| \le \frac{1}{2}\overline{L}\sqrt{n}\delta \le \overline{\vartheta} \triangleq \underline{J}\lambda(2\kappa) \le \gamma \underline{A}.$$

Remark 61 Theorem 51 shows that the difference $\delta > 0$ can be chosen in such a way that the matrix determined by the relation (D) satisfies the condition for global convergence of either a line search or a trust region method. It is clear that the difference δ is proportional to the number $\underline{J}$ in (J4) and inversely proportional to the numbers $\overline{J}$ and $\overline{L}$ in (J3) and (J5). For the line search method ($\overline{\text{S1}}$)-($\overline{\text{S2}}$) we use $\gamma = (1-\overline{\omega})/(1+\overline{\omega})$. For the trust region method ($\overline{\text{T1}}$)-($\overline{\text{T3}}$) we use $\gamma = (1-2\underline{\rho})\underline{\sigma}$.

5.4 Quasi-Newton methods

Definition 34 We say that the basic method for solving systems of nonlinear equations is a quasi-Newton methd if

$$A_i s_i + f_i = 0, \tag{QN1}$$

where A_i, $i \in N$, are nonsingular matrices constructed using the recurrence relation

$$A_{i+1} = A_i + u_i v_i^T, \tag{QN2}$$

where $u_i \in R^n$, $v_i \in R^n$, and satisfying the condition

$$A_{i+1} d_i = y_i, \tag{QN3}$$

where $y_i = f_{i+1} - f_i$, $d_i = x_{i+1} - x_i$.

Remark 62 In this section we deal only with exact quasi-Newton methods (condition (QN1)), so that $(A_i s_i + f_i)/\parallel f_i \parallel = 0\ \forall i \in N$. But $(A_i - J_i)s_i/\parallel f_i \parallel = 0$, $\forall i \in N$, cannot hold (the matrices A_i can substantially differ from the matrices J_i).

Theorem 52 Let $A_+ = A + uv^T$ and $Ad \ne y$. Then $A_+ d = y$ if and only if $v^T d \ne 0$ and $u = (y - Ad)/v^T d$, so that

$$A_+ = A + \frac{(y - Ad)v^T}{v^T d}. \tag{$\overline{\text{A}}$}$$

If $Ad = y$, then it suffices to set $u = v = 0$, so that $A_+ = A$.

Proof From the condition $A_+ d = y$ we obtain $A_+ d = Ad + uv^T d = y$. If $Ad = y$, then it suffices to set $u = v = 0$, so that $A_+ = A$. If $Ad \ne y$, then $v^T d \ne 0$ and $u = (y - Ad)/v^T d$ has to hold.

Remark 63 Setting $v = d$, we obtain the Broyden good update

$$A_+ = A + \frac{(y - Ad)d^T}{d^T d}. \tag{$\overline{\text{AG}}$}$$

Setting $v = A^T y$, we obtain the Broyden bad update

$$A_+ = A + \frac{(y - Ad)y^T A}{y^T A d} \tag{$\overline{\text{AB}}$}.$$

Let

$$e_k^T d = \max_{1 \le i \le n} e_i^T d$$

Setting $v = e_k$, we obtain the column update

$$A_+ = A + \frac{(y - Ad)e_k^T}{e_k^T d}, \tag{$\overline{\mathrm{AD}}$}$$

which changes only one column of the matrix A.

Theorem 53 Let A be a nonsingular matrix and let ($\overline{\mathrm{A}}$) hold. Then the matrix A_+ is nonsingular if and only if $v^T A^{-1} y \neq 0$.

Proof Let $A_+ = A + uv^T$. Then, using the Sherman-Morrison formula, we obtain

$$A_+^{-1} = A^{-1} - \frac{A^{-1} u v^T A^{-1}}{1 + v^T A^{-1} u},$$

so that A_+ is nonsingular if and only if $1 + v^T A^{-1} u \neq 0$. Substituting $u = (y - Ad)/v^T d$ to this inequality, we get

$$1 + v^T A^{-1} u = 1 + \frac{v^T A^{-1} y - v^T d}{v^T d} = \frac{v^T A^{-1} y}{v^T d},$$

so that A_+ is nonsingular if and only if $v^T A^{-1} y \neq 0$.

Remark 64 Theorem 53 gives the reason for the use of Broyden bad method. If $y \neq 0$ and the matrix A is nonsingular, then the choice $v = A^T y$ gives $v^T A^{-1} y = y^T A A^{-1} y = y^T y = \| y \|^2 \neq 0$.

Theorem 54 (Update of the matrix$S = A^{-1}$). Let the assumptions of Theorem 53 be satisfied. Let $S = A^{-1}$ and let A_+ be the matrix determined using the update ($\overline{\mathrm{A}}$), where $v^T A^{-1} y \neq 0$. Let $S_+ = A_+^{-1}$. Then

$$S_+ = S + \frac{(d - Sy) v^T S}{v^T S y}. \tag{$\overline{\mathrm{S}}$}$$

Proof The Sherman-Morrison formula (see proof of Theorem 53) implies

$$S_+ = S - \frac{S u v^T S}{\delta} = S + \frac{(d - Sy) v^T S}{\delta v^T d},$$

where δ is a number to be determined. But the equation $S_+ y = d$ gives

$$S_+ y = Sy + \frac{v^T S y}{\delta v^T d}(d - Sy) = d,$$

so that necessarily $\delta = v^T S y / v^T d$.

Remark 65 Setting $v = d$, we obtain the Broyden good update

$$S_+ = S + \frac{(d - Sy) d^T S}{d^T S y}. \tag{$\overline{\mathrm{SG}}$}$$

Setting $v = (S^{-1})^T y$, we obtain the Broyden bad update

$$S_+ = S + \frac{(d - Sy) y^T}{y^T y}. \tag{$\overline{\mathrm{SB}}$}$$

Let

$$e_k^T y = \max_{1 \le i \le n} e_i^T y.$$

Setting $S^T v = e_k$ we get the inverse column update

$$S_+ = S + \frac{(d - Sy)e_k}{e_k^T y}. \tag{$\overline{\mathrm{SI}}$}$$

Remark 66 (Duality). The relation ($\overline{\mathrm{S}}$) can be obtained from the relation ($\overline{\mathrm{A}}$) by the replacement $d \to y$, $y \to d$, $A \to S$. The good and bad Broyden method are dual to each other. Similarly, the column update and the inverse column update are dual to each other.

Remark 67 For practical use, we recommend the Broyden good method and the column update. The Broyden bad method and the inverse column update (in the dense case) are less efficient.

Quasi-Newton methods for solving systems of nonlinear equations satisfy the quasi-Newton condition similarly as variable metric methods for unconstrained optimization (it suffices to compare (QN3) with (VM3)). Variable metric metods with perfect line searches find a minimum of the quadratic function (Q) after a finite number of steps. We show that quasi-Newton methods with unit stepsizes ($\alpha_i = 1\ \forall i \in N$) find a solution of the linear system

$$J^*(x - x^*) = 0 \tag{L}$$

with the nonsingular matrix J^* also after a finite number of steps. In the proof of this statement, we use the expressions

$$x_{i+1} = x_i - S_i f_i \tag{α}$$

and

$$S_{i+1} = S_i + \frac{(d_i - S_i y_i) z_i^T}{z_i^T y_i} \tag{β}$$

$\forall i \in N$, where S_i are nonsingular matrices, $f_i \neq 0$ and $z_i^T y_i \neq 0\ \forall i \in N$ (here $z_i = S_i^T v_i$).

Lemma 18 Consider the iterative process (α), (β) applied to the linear system (L) with the nonsingular matrix J. Then, for an arbitrary index $i \in N$ and for an arbitrary exponent $k \geq 0$, the vector $(J^* S_{i+1})^k f_{i+1}$ is a linear combination of the vectors $(I - J^* S_i)(J^* S_i)^j f_i$, $0 \leq j \leq k$.

Proof (by induction). Suppose that for some exponent $k \geq 0$ the vector $(J^* S_{i+1})^k f_{i+1}$ is a linear combination of the vectors $(I - J^* S_i)(J^* S_i)^j f_i$, $0 \leq j \leq k$. This obviously holds for $k = 0$, since (L) and (α) imply

$$y_i = f_{i+1} - f_i = J^* d_i = -J^* S_i f_i, \tag{γ}$$

so that

$$(J^* S_{i+1})^0 f_{i+1} = f_{i+1} = f_i + y_i = f_i - J^* S_i f_i = (I - J^* S_i)(J^* S_i)^0 f_i.$$

Using (β) and (γ), we obtain

$$J^* S_{i+1} = J^* S_i + (J^* d_i - J^* S_i y_i) \frac{z_i^T}{z_i^T y_i} = J^* S_i - (I - J^* S_i) J^* S_i f_i \frac{z_i^T}{z_i^T y_i}.$$

since the vector $(J^* S_{i+1})^k f_{i+1}$ is a linear combination of the vectors $(I - J^* S_i)(J^* S_i)^j f_i$, $0 \leq j \leq k$, and since the matrices $J^* S_i$ and $(I - J^* S_i)$ commute, the vector $(J^* S_{i+1})^{k+1} f_{i+1} = J^* S_{i+1} (J^* S_{i+1})^k f_{i+1}$ is a linear combination of the vectors $(I - J^* S_i)(J^* S_i)^j f_i$, $0 \leq j \leq k + 1$.

Lemma 19 Let the assumptions of Lemma 18 be satisfied and let $i \in N$ be an index such that the vector f_{i+1} is not a multiple of the vector f_i. Then the vectors $(J^* S_{i-2l+2})^k f_{i-2l+2}$, $0 \leq k \leq l$, are linearly independent for an arbitrary number $l \in N$, such that $2l \leq i + 1$.

Proof (by induction). Suppose that the vectors $(J^*S_{i-2l+2})^k f_{i-2l+2}$, $0 \le k \le l$, are linearly independent for some integer $l \in N$, such that $2l \le i-1$. This obviously holds for $l=1$, since by (γ) we obtain

$$\begin{aligned} (J^*S_i)^0 f_i &= f_i, \\ (J^*S_i)^1 f_i &= -y_i = f_i - f_{i+1} \end{aligned}$$

and these vectors are linearly independent, since the vector f_{i+1} is not a multiple of the vector f_i.
(a) By Lemma 18, the vector $(J^*S_{i-2l+2})^k f_{i-2l+2}$ is a linear combination of the vectors $(I-J^*S_{i-2l+1})(J^*S_{i-2l+1})^j f_{i-2l+1}$, $0 \le j \le k$. Since the $l+1$ linearly independent vectors $(J^*S_{i-2l+2})^k f_{i-2l+2}$, $0 \le k \le l$, are expressed by using $l+1$ vectors $(I-J^*S_{i-2l+1})(J^*S_{i-2l+1})^k f_{i-2l+1}$, $0 \le k \le l$, then also these vectors have to be linearly independent. This consideration immediately implies that the vectors $(J^*S_{i-2l+1})^k f_{i-2l+1}$, $0 \le k \le l$, are linearly independent.
(b) Using (γ), we obtain

$$y_{i-2l} = -J^*S_{i-2l} f_{i-2l} \ne 0.$$

We show that the vector y_{i-2l} is not a linear combination of the vectors $(J^*S_{i-2l+1})^k f_{i-2l+1}$, $0 \le k \le l$. Using the quasi-Newton condition

$$S_{i-2l+1} y_{i-2l} = d_{i-2l} = (J^*)^{-1} y_{i-2l},$$

we can write

$$(I - J^*S_{i-2l+1}) y_{i-2l} = 0 \qquad (\delta).$$

Suppose that the vector y_{i-2l} is a linear combination of the vectors $(J^*S_{i-2l+1})^k f_{i-2l+1}$, $0 \le k \le l$. Then the corresponding linear combination of the vectors $(I-J^*S_{i-2l+1})(J^*S_{i-2l+1})^k f_{i-2l+1}$, $0 \le k \le l$, has to be zero (see (δ), which is a contradiction to linear independence of these vectors (see (a)).
(c) By Lemma 18, the vector $(J^*S_{i-2l+1})^k f_{i-2l+1}$, is a linear combination of the vectors $(I - J^*S_{i-2l})(J^*S_{i-2l})^j f_{i-2l}$, $0 \le j \le k$, so that it is also a linear combination of the vectors $(J^*S_{i-2l})^j f_{i-2l}$, $0 \le j \le k+1$. Moreover, the vector y_{i-2l} can be expressed in the form $y_{i-2l} = -J^*S_{i-2l} f_{i-2l}$, (see (γ)). Since we express $l+2$ linearly independent vectors y_{i-2l} and $(J^*S_{i-2l+1})^k f_{i-2l+1}$, $0 \le k \le l$ (see (b)) by $l+2$ vectors $(J^*S_{i-2l})^k f_{i-2l}$, $0 \le k \le l+1$, also these vectors have to be linearly independent.

Theorem 55 Let the assumptions of Lemma 18 be satisfied. Then an index $1 \le i \le 2n-1$ exists, such that $f_{i+2} = 0$, so that the point $x_{i+2} \in R^n$ is a solution of the linear system (L).

Proof Suppose that, for $i = 2n-1$, the vector f_{i+1} is not a multiple of the vector f_i. Then, by Lemma 19, the vectors $(J^*S_{2n-2l+1})^k f_{2n-2l+1}$, $0 \le k \le l$, are linearly independent for an arbitrary integer $l \in N$, such that $l \le n$. If $l = n$, then we have $n+1$ linearly independent vectors, which is a contradiction to the fact that they have dimension n. Therefore, an index $1 \le i \le 2n-1$ exists, such that the vector f_{i+1} is a multiple of the vector f_i or

$$f_{i+1} = \lambda_i (f_{i+1} - f_i) = \lambda_i y_i.$$

The expressions (β) and (γ) implies that

$$f_{i+2} = f_{i+1} + y_{i+1} = f_{i+1} - J^*S_{i+1} f_{i+1} = \lambda_i (y_i - J^*S_{i+1} y_i) = \lambda_i (y_i - J^* d_i) = \lambda_i (y_i - y_i) = 0,$$

so that the point $x_{i+2} \in R^n$ is a solution of the linear system (L).

A disadvantage of quasi-Newton methods is the fact that global convergence is not guaranteed (the matrices A_i, $i \in N$, can be a bad approximations of the Jacobian matrix J_i, $i \in N$ in general). Therefore, these methods have to be combined with the discrete Newton method. Line search quasi-Newton methods are usually realized in such a way that $A_1 = J_1$. Whenever it is not possible to satisfy the condition ($\overline{\text{S2a}}$) (or ($\overline{\text{S2b}}$), or ($\overline{\text{S2c}}$)), the iterative process is restarted (we set $A_{i+1} = J_{i+1}$).

Trust region quasi-Newton methods are usually realized in such a way that $A_1 = J_1$. In the case ($\overline{\text{T3a}}$), we set $A_{i+1} = J_{i+1}$. In the case ($\overline{\text{T3b}}$), the matrix A_{i+1} is obtained by the update ($\overline{\text{A}}$). These modifications have a theoretical support, since the following proposition holds.

Proposition 9 Let $x^* \in R^n$ be a point, such that $f(x^*) = 0$ and the matrix $J(x^*)$ is nonsingular. Then numbers $\overline{\delta} > 0$ and $\overline{\vartheta} > 0$ exist, such that if $\| x_1 - x^* \| \le \overline{\delta}$ and $\| A_1 - J_1 \| \le \overline{\vartheta}$, then the sequence x_i, $i \in N$, determined by Broyden good method ($\overline{\text{AG}}$) with unit stepsizes ($\alpha_i = 1 \; \forall i \in N$) converges Q-superlinearly to the point $x^* \in R^n$.

Proposition 9 is a special case of Theorem 57.

The following table shows a comparison of the discrete Newton method with the Broyden good method for solving 28 test systems of nonlinear equations with 2-16 unknowns (the total number of iterations NIT, function evaluations NFV and numerical differentiations NFJ, is shown together with the total computational time). Both methods were realized as trust region methods.

Method	NIT-NFV-NFJ	time
discrete Newton	504-5890-504	6.37
good Broyden	723-1844-93	2.75

6 Methods for large-scale sparse systems of nonlinear equations

Large-scale systems of nonlinear equations cannot be solved by methods which require dense matrices. Some special methods have to used in the large-scale case.

- Quasi-Newton methods with limited storage.
- Difference version of the inexact Newton method.
- Sparse discrete Newton method.
- Sparse quasi-Newton methods.
- Methods that use special updates.

6.1 Quasi-Newton methods with limited storage

Quasi-Newton methods with limited storage are based on the use of a limited number of steps of the Broyden good method or the column update. Let $M = \{i \in N : \; i = (j-1)m + 1, \; j \in N\}$, where m is a number of steps of a quasi-Newton method with limited storage. Then we set $S_l = (J_l^{-1})$ for $l \in M$ and

$$S_{i+1} = S_i + \frac{(d_i - S_i y_i) v_i^T S_i}{v_i^T S_i y_i} = (I + w_i v_i^T) S_i$$

for $l \le i \le l + m$ (see ($\overline{\text{S}}$)), where $v_i = d_i$ (the Broyden good method) or $v_i = e_k$ (the column update) and

$$w_i = \frac{d_i - S_i y_i}{v_i^T S_i y_i}.$$

The vectors $v_i \in R^n$, $w_i \in R^n$, $l \le i \le l + m$, are stored.

Knowing the vectors $v_j \in R^n$, $w_j \in R^n$, $l \le j \le l+m$, we first determine the vector $p_l^{i+1} = -S_l f_{i+1}$ (the matrix S_l is usually represented by the triangular decomposition $(S_l)^{-1} = L_l U_l$, which is either a complete or incomplete LU decomposition of the matrix J_l). Then we compute the vectors

$$p_{j+1}^{i+1} = (I + w_j v_j^T) p_j^{i+1}$$

for $l \le j \le i-1$. Finally we determine the vectors v_i and

$$w_i = \frac{d_i - (p_i^{i+1} + s_i)}{v_i^T (p_i^{i+1} + s_i)},$$

where $s_i = -S_i f_i$ is the direction vector from the previous iteration (usually $s_i = d_i/\alpha_i$) and set

$$s_{i+1} = -S_{i+1} f_{i+1} = -(I + w_i v_i^T) p_i^{i+1}.$$

Quasi-Newton methods with limited storage can be also realized using a compact form. For derivation of this compact form, we use the notation $D_k = [d_1, \ldots, d_k]$, $Y_k = [y_1, \ldots, y_k]$, $V_k = [v_1, \ldots, v_k]$. Furthermore, we denote R_k the upper triangular matrix of order k, such that $(R_k)_{ij} = v_i^T d_j$, $i \le j$ a $(R_k)_{ij} = 0$, $i > j$. To simplify the notation in the subsequent proofs, we omit the index k and replace the index $k+1$ by $+$. For this convenience, we denote $D = [d_1, \ldots, d_{k-1}]$, $Y = [y_1, \ldots, y_{k-1}]$, $V = [v_1, \ldots, v_{k-1}]$ and $R = R_{k-1}$, so that $D_k = [D, d]$, $Y_k = [Y, y]$, $V_k = [V, v]$ and

$$R_k = \begin{bmatrix} R, & V^T d \\ 0, & v^T d \end{bmatrix}.$$

Theorem 56 Let A_1 be a nonsingular matrix and let ($\overline{\text{A}}$) hold with $v_k^T d_k \ne 0$ for an arbitrary index $1 \le k \le m$. Then we can write

$$A_{k+1} = A_1 + (Y_k - A_1 D_k) R_k^{-1} V_k^T. \tag{$\overline{\text{AA}}$}$$

Proof For $k = 1$, the expression ($\overline{\text{AA}}$) is equivalent to ($\overline{\text{A}}$). The induction step: Suppose that ($\overline{\text{AA}}$) holds for all indices less than k. At the stage k, the formula ($\overline{\text{AA}}$) can be written in the form

$$A_+ = A_1 + [Y - A_1 D, y - A_1 d] \begin{bmatrix} R, & V^T d \\ 0, & v^T d \end{bmatrix}^{-1} \begin{bmatrix} V^T \\ v^T \end{bmatrix}.$$

Since

$$\begin{bmatrix} R, & V^T d \\ 0, & v^T d \end{bmatrix}^{-1} = \begin{bmatrix} R^{-1}, & -\frac{R^{-1} V^T d}{v^T d} \\ 0, & \frac{1}{v^T d} \end{bmatrix}$$

(which can be easily verified by multiplication), we can write

$$A_+ = A_1 + (Y - A_1 D) R^{-1} V^T \left(I - \frac{d v^T}{v^T d} \right) + (y - A_1 d) \frac{v^T}{v^T d} = A + \frac{(y - Ad) v^T}{v^T d},$$

which is exactly the relation ($\overline{A}$).

Remark 68 Direct inversion of the relation ($\overline{AA}$) (using the Woodbury formula) gives

$$A_{k+1}^{-1} = A_1^{-1} - A_1^{-1}(Y_k - A_1 D_k)(R_k + V_k^T A_1^{-1}(Y_k - A_1 D_k))^{-1} V_k^T A_1^{-1}$$

or

$$S_{k+1} = S_1 + (D_k - S_1 Y_k)(C_k - L_k + V_k^T S_1 Y_k)^{-1} V_k^T S_1, \tag{$\overline{\text{SS}}$}$$

where L_k is the lower triangular matrix, such that $(L_k)_{ij} = 0$, $i < j$, and $(L_k)_{ij} = v_i^T d_j$, $i \ge j$, and C_k is the diagonal matrix of order k, such that $(C_k)_{ij} = v_i^T d_j$, $i = j$, and $(C_k)_{ij} = 0$, $i \ne j$.

The compact form is most frequently used in connection with iterative solution of the linear system $A_i s_i + f_i = 0$, $i \in N$. We set $A_l = J_l$ for $l \in N$ and

$$A_{i+1} = A_l + (Y_k - A_l D_k) R_k^{-1} V_k^T$$

for $l \le i \le l + m$ (see ($\overline{\text{AA}}$), where $D_k = [d_l, \ldots, d_i]$, $Y_k = [y_l, \ldots, y_i]$, $V_k = [v_l, \ldots, v_i]$ and R_k is the upper triangular matrix of order $k = i - l + 1$, such that $(R_k)_{ij} = v_{l+i-1}^T d_{l+j-1}$, $i \le j$, and $(R_k)_{ij} = 0$,

$i > j$. Notice that the matrix V_k is not usually stored (the Broyden good method uses $V_k = D_k$ and the column update only suffices to store indices of the maximum absolute value elements of columns of the matrix D_k). Instead of the matrix Y_k, we store the matrix $U_k = Y_k - A_l D_k$ and the product $A_{i+1}p$ is computed by the formula $A_{i+1}p = A_l p + U_k R_k^{-1} V_k^T p$.

6.2 Difference version of the inexact Newton method

Difference version of the inexact Newton method are based on the inexact iterative solution of the linear system $Js + f = 0$. The matrix J is not used since the multiplication $q = Jp$ is replaced by the numerical differentiation

$$J(x)p \approx \frac{f(x+\delta p) - f(x)}{\delta},$$

where δ is a small difference ($\delta = \sqrt{\varepsilon_M}/ \parallel p \parallel$ and ε_M is a machine precision). If the calculation of the vector $f(x)$ requires $O(n)$ operations, then this approach can be more efficient than matrix by vector multiplication ($O(n^2)$ operations in general). Moreover, the Jacobian matrix has to be neither computed nor stored. However, methods for iterative solution of the linear system $Js + f = 0$ have not to use the transposed Jacobian matrix, which slightly restricts their selection (iterative methods for solving systems of linear equations are described in Section 6.7).

6.3 Sparse discrete Newton method

The sparse discrete Newton method can be realized in two different forms which differ by organization of the numerical differentiation (elementwise or groupwise differentiation). Elementwise differentiation determines individual nonzero elements of the Jacobian matrix by the formulas

$$(J(x))_{ij} \approx \frac{f_i(x+\delta e_j) - f_i(x)}{\delta}.$$

Only the differences $f_i(x+\delta e_j) - f_i(x)$ which correspond to nonzero elements $(J(x))_{ij} \neq 0$ are computed.

Groupwise differentiation is based on the approximation of columns Je_j, $1 \leq j \leq n$, of the Jacobian matrix J by the formula

$$J(x)e_j \approx \frac{f(x+\delta e_j) - f(x)}{\delta},$$

where δ is a small difference ($\varepsilon = \sqrt{\varepsilon_M}$). However, if the matrix J is sparse, then usually one difference of the function vector allows us to determine several columns of J (similarly to Section 4.3). Divide columns of the matrix J into k disjoint groups $\mathcal{S}_i \subset \{1, \ldots, n\}$, $1 \leq i \leq k$, so that each submatrix $J(\mathcal{S}_i)$, consisting of columns of the matrix J belonging to the group $\mathcal{S}_i$, has at most one nonzero element in each row. Then all nonzero elements of the matrix J can be determined using k differences

$$\frac{f(x+\delta v_i) - f(x)}{\delta} \approx Jv_i \qquad 1 \leq i \leq k,$$

where v_i, $1 \leq i \leq k$, are vectors consisting only of zeroes and units, such that

$$(v_i)_j = e_j^T v_i = 1 \Longleftrightarrow j \in \mathcal{S}_i$$

(using the vector v_i we determine the submatrix $J(\mathcal{S}_i)$). The partition $\{1, \ldots, n\} = \mathcal{S}_i \cup \ldots \cup \mathcal{S}_k$, is determined in such a way that the number of groups k is as small as possible. Since the number of groups cannot be less then the number of nonzero elements in an arbitrary row, the number of function vector evaluations is usually slightly greater than the number given by elementwise differentiation. On the other hand, the computation can now be organized better (the expressions which are common for all scalar functions can be computed only k times), so that the groupwise differentiation is usually faster.

Using the sparse discrete Newton method is supported by the theory introduced in Section 5.3 (Lemma 17).

6.4 Sparse quasi-Newton methods

Sparse quasi-Newton methods are based on updates which preserve the sparsity pattern of the Jacobian matrix. Denote

$$\begin{aligned}\mathcal{V}_Q &= \{A \in R^{n\times n} : Ad = y\},\\ \mathcal{V}_G &= \{A \in R^{n\times n} : J_{ij} = 0 \Rightarrow A_{ij} = 0\}.\end{aligned}$$

In the same way as in Section 4.4, we can define the orthogonal projections $\mathcal{P}_Q$, $\mathcal{P}_G$ into the linear manifolds $\mathcal{V}_Q$, $\mathcal{V}_G$ by the formulas

$$\begin{aligned}\mathcal{P}_Q A &= \min_{A_+\in\mathcal{V}_Q} \| A_+ - A \|_F,\\ \mathcal{P}_G A &= \min_{A_+\in\mathcal{V}_G} \| A_+ - A \|_F,\end{aligned}$$

Similarly we can define the orthogonal projection $\mathcal{P}_{QG}$ into $\mathcal{V}_Q \cap \mathcal{V}_G$. Using Theorem 36, we obtain

$$\mathcal{P}_{QG} A = \mathcal{P}_G(A + ud^T)$$

where the vector $u \in R^n$ is a solution of the system $Qu = y - Ad$ with the diagonal positive semidefinite matrix

$$Q = \sum_{i=1}^{n} \| d^i \|^2 \, e_i e_i^T,$$

where d^i, $1 \le i \le n$, are vectors, such that $d^i_j = d_j$, $J_{ij} \neq 0$ and $d^i_j = 0$, $J_{ij} = 0$. Denoting $A_+ = \mathcal{P}_{QG}A$, we can formally write the formula $\mathcal{P}_{QG}A = \mathcal{P}_G(A + ud^T)$ in the form

$$A_+ = A + \sum_{i=1}^{n} \frac{e_i^T(y - Ad)e_i(d^i)^T}{(d^i)^T d^i}, \tag{$\overline{\text{AS}}$}$$

where terms with $d^i = 0$ are omitted. The method which utilizes the update ($\overline{\text{AS}}$) is called the Schubert method and since it is a generalization of the Broyden good method, it has similar properties as the Broyden good method. Since global convergence of the Schubert method is not guaranteed, it is necessary to use restarts setting $A_+ = J_+$. However, it is possible to prove that the Schubert method converges locally Q-superlinearly.

Lemma 20 Let A_+ be the matrix determined by ($\overline{\text{AS}}$). Then for an arbitrary matrix $\tilde{J} \in \mathcal{V}_Q \cap \mathcal{V}_G$ it holds

$$\| A_+ - \tilde{J} \|_F^2 \le \| A - \tilde{J} \|_F^2 - \frac{\| y - Ad \|^2}{\| d \|^2}.$$

Proof Since $\tilde{J} \in \mathcal{V}_Q \cap \mathcal{V}_G$, $\mathcal{P}_{QG}$ is the orthogonal projection into $\mathcal{V}_Q \cap \mathcal{V}_G$ and since $A_+ = \mathcal{P}_{QG}A$, we can use Pythagoras theorem

$$\| A_+ - \tilde{J} \|_F^2 = \| A - \tilde{J} \|_F^2 - \| A_+ - A \|_F^2 .$$

Since $\mathcal{V}_Q \cap \mathcal{V}_G \subset \mathcal{V}_Q$ implies $A_+ d = y$, we can write

$$\| y - Ad \| = \| (A_+ - A)d \| \le \| A_+ - A \| \| d \| \le \| A_+ - A \|_F \| d \| .$$

which after substitution gives the statement of the lemma.

Lemma 21 Let A_+ be the matrix determined by ($\overline{\text{AS}}$) and let (J5) hold. Then

$$\| A_+ - J_+ \|_F \le \| A - J \|_F + \overline{L}\sqrt{n} \, \| d \| .$$

Proof Denote

$$\tilde{J} = \int_0^1 J(x + \lambda d) d\lambda.$$

By the same way as in the part (a) of the proof of Theorem 38 (using the mean value theorem), we can show that

$$\begin{aligned} \| \tilde{J} - J \|_F &\leq \frac{1}{2} \overline{L} \sqrt{n} \| d \|, \\ \| \tilde{J} - J_+ \|_F &\leq \frac{1}{2} \overline{L} \sqrt{n} \| d \| . \end{aligned}$$

Using Lemma 20, we get

$$\begin{aligned} \| A_+ - J_+ \|_F &\leq \| A_+ - \tilde{J} \|_F + \| \tilde{J} - J_+ \|_F \leq \| A - \tilde{J} \|_F + \| \tilde{J} - J_+ \|_F \leq \\ &\leq \| A - J \|_F + \| \tilde{J} - J \|_F + \| \tilde{J} - J_+ \|_F, \end{aligned}$$

which after substitution gives the statement of the lemma.

Theorem 57 Let the condition (J5) be satisfied and let $x^* \in R^n$ be a point, such that $f(x^*) = 0$ and $J(x^*)$ is nonsingular. Then numbers $\overline{\delta} > 0$, $\overline{\lambda} > 0$ exist, such that if $\| x_1 - x^* \| \leq \overline{\delta}$, $\| A_1 - J_1 \| \leq \overline{\lambda}$ and if it holds

$$\begin{aligned} \| A_i d_i + f_i \| &\leq \overline{\omega} \| f_i \| \\ x_{i+1} &= x_i + d_i \\ A_{i+1} &= \mathcal{P}_{QG} A_i \end{aligned}$$

$\forall i \in N$, where $0 \leq \overline{\omega} < 1$ (inexact Schubert method), then the sequence x_i, $i \in N$, converges to the point $x^* \in R^n$. If in addition $\| \omega_i \| = \| A_i d_i + f_i \| / \| f_i \| \to 0$, then $x_i \to x^*$ Q-superlinearly.

Proof Results obtained in parts (a) - (b) of the proof of Theorem 47 can be reformulated (using the concept of neighborhood) to prove that numbers $\delta > 0$, $\vartheta > 0$ exist, such that if $\| x - x^* \| \leq \delta$, $\| (A - J(x)) d \| \leq \vartheta \| d \|$ and $\| Ad + f \| \leq \overline{\omega} \| f \|$, where $0 \leq \overline{\omega} < 1$, then it holds

$$\frac{1 - \overline{\omega}}{\overline{J}} \| f \| \leq \| d \| \leq \frac{1 + \overline{\omega}}{\underline{J}} \| f \|,$$

where $\| J^* \| < \overline{J}$, $\| (J^*)^{-1} \| < 1/\underline{J}$ and

$$\| f(x + d) \| \leq r \| f \|$$

(where $\overline{\omega} < r < 1$). Notice that the number $0 \leq \overline{\omega} < 1$ can be arbitrary, while the numbers $\delta > 0$ and $\vartheta > 0$ can be required to be very small.
(a) Choose the numbers $\overline{\delta} > 0$ and $\overline{\vartheta} > 0$ so that

$$\overline{\delta} \left(1 + \frac{\overline{J}}{\underline{J}} \frac{1 + \overline{\omega}}{1 - r} \right) \leq \delta$$

and

$$\overline{\vartheta} + \overline{L} \frac{\overline{J}}{\underline{J}} \frac{1 + \overline{\omega}}{1 - r} \overline{\delta} \leq \vartheta / \sqrt{n}.$$

Let $\| x_1 - x^* \| \leq \overline{\delta}$ and $\| (A_1 - J(x_1) \| \leq \overline{\vartheta}$. We prove by induction that for an arbitrary index $i \in N$ $\| x_i - x^* \| \leq \delta$ and $\| A_i - J(x_i) \| \leq \vartheta$ hold. For $i = 1$, it is obvious. Assume validity of this statement for $1 \leq i \leq k$. Then

$$\begin{aligned}
\| x_{k+1} - x^* \| &\le \| x_1 - x^* \| + \sum_{i=1}^{k} \| d_i \| \le \| x_1 - x^* \| + \frac{1+\overline{\omega}}{\underline{J}} \sum_{i=1}^{k} \| f_i \| \le \\
&\le \| x_1 - x^* \| + \frac{1+\overline{\omega}}{\underline{J}} \| f_1 \| \sum_{i=1}^{k} r^{i-1} \le \| x_1 - x^* \| + \frac{\overline{J}}{\underline{J}} \frac{1+\overline{\omega}}{1-r} \| x_1 - x^* \| \le \\
&\le \overline{\delta} \left(1 + \frac{\overline{J}}{\underline{J}} \frac{1+\overline{\omega}}{1-r} \right) \le \delta
\end{aligned}$$

and using Lemma 20, we obtain

$$\begin{aligned}
\frac{\| (A_{k+1} - J_{k+1}) d_{k+1} \|}{\| d_{k+1} \|} &\le \| A_{k+1} - J_{k+1} \| \le \| A_{k+1} - J_{k+1} \|_F \le \\
&\le \| A_1 - J_1 \|_F + \overline{L}\sqrt{n} \sum_{i=1}^{k} \| d_i \| \le \overline{\vartheta}\sqrt{n} + \overline{L}\sqrt{n} \frac{\overline{J}}{\underline{J}} \frac{1+\overline{\omega}}{1-r} \overline{\delta} \le \vartheta.
\end{aligned}$$

(b) By (a) it holds $\| f_{i+1} \| \le r \| f_i \| \le r^i \| f_1 \| \; \forall i \in N$, where $\overline{\omega} < r < 1$, so that $\sum_{i=1}^{\infty} \|f_i\| < \infty$, $\sum_{i=1}^{\infty} \|d_i\| < \infty$ and, therefore, also $\| f_i \| \to 0$, $\| d_i \| \to 0$ and $x_i \to x^*$.
(c) From Lemma 20 it follows

$$\begin{aligned}
\frac{\| y - Ad \|^2}{\| d \|^2} &\le \| A - \tilde{J} \|_F^2 - \| A_+ - \tilde{J} \|_F^2 = \\
&= \left(\| A - \tilde{J} \|_F - \| A_+ - \tilde{J} \|_F \right) \left(\| A - \tilde{J} \|_F + \| A_+ - \tilde{J} \|_F \right) \le \\
&\le \overline{M} \left(\| A - \tilde{J} \|_F - \| A_+ - \tilde{J} \|_F \right).
\end{aligned}$$

Existence of the constant $\overline{M}$ follows from the fact that

$$\begin{aligned}
\| A - \tilde{J} \|_F + \| A_+ - \tilde{J} \|_F &\le \| A - J \|_F + \| A_+ - J_+ \|_F + \overline{L}\sqrt{n} \| d \| \le \\
&\le 2 \| A - J \|_F + 2\overline{L}\sqrt{n} \| d \| \le \\
&\le 2 \| A - J \|_F + 2\overline{L}\sqrt{n} \left(\| x^+ - x^* \| + \| x - x^* \| \right),
\end{aligned}$$

so that (a) implies

$$\| A - \tilde{J} \|_F + \| A_+ - \tilde{J} \|_F \le 2\sqrt{n}\vartheta + 4\overline{L}\sqrt{n}\delta \triangleq \overline{M}.$$

Furthermore, we can write

$$\| A_+ - J_+ \|_F \le \| A_+ - \tilde{J} \|_F + \| J_+ - \tilde{J} \|_F,$$

so that

$$\begin{aligned}
\| A - \tilde{J} \|_F - \| A_+ - \tilde{J} \|_F &\le \| A - J \|_F + \| J - \tilde{J} \|_F - \| A_+ - J_+ \|_F + \| J_+ - \tilde{J} \|_F \le \\
&\le \| A - J \|_F - \| A_+ - J_+ \|_F + \overline{L}\sqrt{n} \| d \|,
\end{aligned}$$

which gives

$$\begin{aligned}
\sum_{i=1}^{\infty} \frac{\| y_i - A_i d_i \|^2}{\| d_i \|^2} &\le \overline{M} \left(\| A_1 - J_1 \|_F - \lim_{i \to \infty} \| A_{i+1} - J_{i+1} \|_F \right) + \overline{ML}\sqrt{n} \sum_{i=1}^{\infty} \| d_i \| \le \\
&\le \overline{M} \| A_1 - J_1 \|_F + \overline{ML} \frac{\overline{J}}{\underline{J}} \frac{1+\overline{\omega}}{1-r} \| x_1 - x^* \| < \infty.
\end{aligned}$$

Therefore, $\| y_i - A_i d_i \| / \| d_i \| \to 0$ necessarily holds, which together with $\| \omega_i \| = \| A_i d_i + f_i \| / \| f_i \| \to 0$ (similarly as in the proof of Theorem 39) implies that $x_i \to x^*$ Q-superlinearly

6.5 Methods based on updates of the triangular decomposition

The system of linear equations $As+f=0$ can be solved either directly or iteratively. Direct solution is based on the use of the nonsymmetric triangular (or LU) decomposition

$$PA = LU$$

where P is a permutation matrix, which is chosen in such a way that the number of new nonzero elements (fill-in) is as small as possible, L is a lower triangular matrix with units on the diagonal and U is an upper triangular matrix. Determination of the permutation matrix P and the sparsity pattern of the triangular matrices L and U is called symbolic decomposition. In contrast of the sparse Choleski decomposition (Section 4.3) it does not suffice to carry out the symbolic decomposition only once (before the iterative process is started), since the row permutations (pivoting) can influence a stability of the elimination process. Therefore, the LU decomposition is a rather complicated and time consuming process, so that it is advantageous to restrict its use. This idea is a basis of methods based on LU decomposition. In contrast to the Schubert method, where the matrix A^+ is chosen to satisfy the quasi-Newton condition $A_+d = y$, $d = x_+ - x$, $y = f_+ - f$, the LU decomposition based methods set $PA_+ = LU_+$ and the matrix U_+ is chosen to satisfy the quasi-Newton condition

$$U_+d = v \triangleq L^{-1}Py.$$

Since the sparsity pattern of the matrix U has to be preserved, we can use the way described in Section 6.4. As a result we obtain the update

$$U_+ = U + \sum_{i=1}^{n} \frac{e_i(v - Ud)e_i d^i}{(d^i)^T d^i}, \tag{$\overline{\mathrm{AD}}$}$$

where d^i, $1 \le i \le n$, are vectors, such that $d^i_j = d_j$, $U_{ij} \ne 0$ and $d^i_j = 0$, $U_{ij} = 0$ (the terms with $d^i = 0$ are omitted). The method which utilizes the update ($\overline{\mathrm{AD}}$) is called the Dennis-Marwil method. This method is usually realized, so that the LU decomposition $PJ = LU$ is carried out first and then, in the m subsequent iterations, we use the update ($\overline{\mathrm{AD}}$). After m updates ($\overline{\mathrm{AD}}$) or after a forced restart, the LU decomposition $PJ = LU$ is carried out again and whole process continues.

The column scaling method is still simpler. In this case, we set $PA_+ = D_+LU$ and the diagonal matrix D_+ is chosen to satisfy the quasi-Newton condition

$$D_+LUd = Py.$$

Writing this condition in the form

$$\sum_{i=1}^{n} D_+e_ie_i^T LUd = Py$$

and utilizing the fact that the matrix D_+ is diagonal, we can write

$$e_i^T D_+e_ie_i^T LUd = e_i^T Py$$

$1 \le i \le n$, or

$$e_i^T D_+e_i = \frac{e_i^T Py}{e_i^T LUd}. \tag{$\overline{\mathrm{AR}}$}$$

Also the row scaling method has to be restarted after m iterations with the determination of the new LU decomposition $PJ = LU$.

6.6 Imperfect versions of the discrete Newton method

Imperfect versions of the discrete Newton method are based on the idea that the numerical differentiation is used only in selected iterations. The simplest one is the Shamanski method, where we set $A = J$ and use the same matrix ($A_+ = A$) in the m subsequent iterations. More sophisticated methods are based on a principle similar to groupwise differentiation. Again a partition $\{1,\dots,n\} = \mathcal{S}_1 \cup \ldots \cup \mathcal{S}_k$

of columns of the matrix J into k disjoing groups $\mathcal{S}_i$, $1 \le i \le k$, is determined, so that each submatrix $J(\mathcal{S}_i)$, consisting of columns of the matrix J belonging to the group $\mathcal{S}_i$, has at most one nonzero element in each row (Section 6.3). Then only one properly chosen group is considered in the current iteration. Columns belonging to this group are determined by numerical differentiation and other columns of the matrix J remain unchanged. More precisely, let $l = \mathrm{mod}_k i$ ($\mathrm{mod}_k i$ is the remainder after division of the numbers i and k). In the i-th iteration we use a vector v_i, such that

$$(v_i)_j = e_j^T v_i = 1 \Longleftrightarrow j \in \mathcal{S}_l$$

and using the difference

$$\frac{f(x + \delta v_i) - f(x)}{\delta} \approx J v_i,$$

we determine the columns of the matrix J belonging to the group $\mathcal{S}_l$. Columns belonging to other groups remain unchanged.

This method, which is called the Li method, can be combined with the Schubert method in such a way that the determination of columns of the matrix J belonging to the group $\mathcal{S}_l$ is followed by the update ($\overline{\text{AS}}$).

6.7 Iterative solution of nonsymmetric linear systems

For solution of the linear system $As + f = 0$ with nonsymmetric matrix A, many iterative methods exist. Roughly speaking, we can divide these methods into two different classes:

- Methods with short recurrences.
- Methods with long recurrences.

An advantage of methods with short recurrences (they are two-term or three-term recurrences) consists in a low number of operations and stored elements, which are $O(n)$. Disadvantage of these methods is a possibility of breakdown (division by zero) during the iterative process. The methods with long recurrences have opposite properties. In the i-th iteration we work with i vectors of dimension n, which require $O(i^2)$ operations and stored elements (theoretically we need n iterations for obtaining the solution which require $O(n^2)$ operations and stored elements). However, breakdowns during iterative process do not appear (each iteration is correctly defined).

In this text, which is only an introduction into iterative methods for solving systems of linear equations, we consider only the smoothed CGS method using short recurrences and the GMRES method using long recurrences.

Definition 35 Let $A \in R^{n\times n}$ be a nonsingular matrix and $f \in R^n$, $\tilde{f} \in R^n$. Then the iterative process described by the recurrence relations

$$s_1 = 0, \qquad f_1 = f, \qquad \tilde{f}_1 = \tilde{f} \qquad p_1 = -f_1, \qquad \tilde{p}_1 = -\tilde{f}_1$$

and

$$q_i = Ap_i, \qquad \tilde{q}_i = A^T \tilde{p}_i, \qquad \alpha_i = \tilde{f}_i^T f_i / \tilde{p}_i^T q_i,$$

$$s_{i+1} = s_i + \alpha_i p_i,$$

$$f_{i+1} = f_i + \alpha_i q_i, \qquad \tilde{f}_{i+1} = \tilde{f}_i + \alpha_i \tilde{q}_i, \qquad \beta_i = \tilde{f}_{i+1}^T f_{i+1} / \tilde{f}_i^T f_i,$$

$$p_{i+1} = -f_{i+1} + \beta_i p_i, \qquad \tilde{p}_{i+1} = -\tilde{f}_{i+1} + \beta_i \tilde{p}_i,$$

for $1 \le i \le n$, is the biconjugate gradient (BCG) method defined by the matrix $A \in R^{n\times n}$ and the vectors $f \in R^n$, $\tilde{f} \in R^n$.

Theorem 58 Consider the biconjugate gradient method defined by a nonsingular matrix $A \in R^{n\times n}$ and vectors $f \in R^n$, $\tilde{f} \in R^n$. Let $\tilde{f}_i f_i \neq 0$ and $\tilde{p}_i q_i \neq 0$ $\forall 1 \leq i \leq n$. Then $f_{n+1} = 0$ and the vector s_{n+1} is a solution of the linear system $As + f = 0$.

Proof Suppose that $\tilde{f}_i^T f_i \neq 0$ and $\tilde{p}_i^T q_i \neq 0$, $\forall 1 \leq i \leq n$. We prove by induction that

$$(\alpha) \qquad \tilde{p}_j^T f_i = p_j^T \tilde{f}_i = 0 \quad \forall 1 \leq j < i \leq n+1,$$

$$(\beta) \qquad \tilde{f}_j^T f_i = f_j^T \tilde{f}_i = 0 \quad \forall 1 \leq j < i \leq n+1,$$

$$(\gamma) \qquad \tilde{p}_j^T q_i = p_j^T \tilde{q}_i = 0 \quad \forall 1 \leq j < i \leq n.$$

From (β) it follows that the vectors f_i, $1 \leq i \leq n$ (and also $\tilde{f}_i$, $1 \leq i \leq n$), are linearly independent. Indeed, if $\lambda_1 f_1 + \ldots + \lambda_n f_n = 0$, then for $1 \leq i \leq n$ it holds

$$\tilde{f}_i^T \left(\sum_{j=1}^{n} \lambda_j f_j \right) = \lambda_i \tilde{f}_i^T f_i = 0$$

and since $\tilde{f}_i^T f_i \neq 0$, we obtain $\lambda_i = 0$. Similarly (γ) implies that the vectors p_i, $1 \leq i \leq n$ (and also $\tilde{p}_i$, $1 \leq i \leq n$), are linearly independent. Since $f_{n+1} = As_{n+1} + f$ (it follows from recurrence relations of the BCG method), the vectors $\tilde{f}_i$, $1 \leq i \leq n$, are linearly independent and

$$\tilde{f}_j^T f_{n+1} = 0 \quad \forall 1 \leq j \leq n,$$

$f_{n+1} = As_{n+1} + f = 0$ has to hold.

For $i = 1$, $(\alpha) - (\gamma)$ holds, since nothing else has to be proved. The induction steps:
(a) Let $i \leq n$. Inductive assumptions (α) and (γ) imply

$$\tilde{p}_j^T f_{i+1} = \tilde{p}_j^T f_i + \alpha_i \tilde{p}_j^T q_i = 0,$$

$$p_j^T \tilde{f}_{i+1} = p_j^T \tilde{f}_i + \alpha_i p_j^T \tilde{q}_i = 0,$$

$\forall 1 \leq j < i$. From (α) and (γ) it follows

$$\tilde{p}_i^T f_{i+1} = \tilde{p}_i^T f_i + \alpha_i \tilde{p}_i^T q_i = -\tilde{f}_i^T f_i + \beta_{i-1} \tilde{p}_{i-1}^T f_i + \frac{\tilde{f}_i^T f_i}{\tilde{p}_i^T q_i} \tilde{p}_i^T q_i = 0,$$

$$p_i^T \tilde{f}_{i+1} = p_i^T \tilde{f}_i + \alpha_i p_i^T \tilde{q}_i = -f_i^T \tilde{f}_i + \beta_{i-1} p_{i-1}^T \tilde{f}_i + \frac{f_i^T \tilde{f}_i}{p_i^T \tilde{q}_i} p_i^T \tilde{q}_i = 0.$$

Therefore, $\tilde{p}_j^T f_{i+1} = 0$, $p_j^T \tilde{f}_{i+1} = 0$ $\forall 1 \leq j \leq i$.
(b) Let $i \leq n$. Recurrence relations of the BCG method imply

$$\tilde{f}_1 = -\tilde{p}_1,$$

$$\tilde{f}_j = -\tilde{p}_j + \beta_{j-1} \tilde{p}_{j-1} \quad \forall 1 < j \leq i,$$

$$f_1 = -p_1,$$

$$f_j = -p_j + \beta_{j-1} p_{j-1} \quad \forall 1 < j \leq i,$$

so that using (a) we obtain

$$\tilde{f}_1^T f_{i+1} = -\tilde{p}_1^T f_{i+1} = 0,$$

$$\tilde{f}_j^T f_{i+1} = -\tilde{p}_j^T f_{i+1} + \beta_{j-1}\tilde{p}_{j-1} f_{i+1} = 0 \quad \forall 1 < j \le i,$$

$$f_1^T \tilde{f}_{i+1} = -p_1^T \tilde{f}_{i+1} = 0,$$

$$f_j^T \tilde{f}_{i+1} = -p_j^T \tilde{f}_{i+1} + \beta_{j-1} p_{j-1} \tilde{f}_{i+1} = 0 \quad \forall 1 < j \le i.$$

(c) Let $i < n$. Recurrence relations of the BCG method and (a) imply

$$\begin{aligned}
\tilde{p}_j^T q_{i+1} &= \tilde{p}_j^T A p_{i+1} = -\tilde{p}_j^T A f_{i+1} + \beta_i \tilde{p}_j^T A p_i = \\
&= -\left(\tilde{f}_{j+1} - \tilde{f}_j\right)^T f_{i+1}/\alpha_j + \beta_i \tilde{p}_j^T q_i = 0, \\
p_j^T \tilde{q}_{i+1} &= p_j^T A^T \tilde{p}_{i+1} = -p_j^T A^T \tilde{f}_{i+1} + \beta_i p_j^T A^T \tilde{p}_i = \\
&= -\left(f_{j+1} - f_j\right)^T \tilde{f}_{i+1}/\alpha_j + \beta_i p_j^T \tilde{q}_i = 0,
\end{aligned}$$

$\forall 1 \le j < i$. Using in addition (b), we obtain

$$\begin{aligned}
\tilde{p}_i^T q_{i+1} &= -\frac{1}{\alpha_i}\left(\tilde{f}_{i+1} - \tilde{f}_i\right)^T f_{i+1} + \beta_i \tilde{p}_i^T q_i = -\frac{\tilde{p}_i^T q_i}{\tilde{f}_i^T f_i}\tilde{f}_{i+1}^T f_{i+1} + \frac{\tilde{f}_{i+1}^T f_{i+1}}{\tilde{f}_i^T f_i}\tilde{p}_i^T q_i = 0, \\
p_i^T \tilde{q}_{i+1} &= -\frac{1}{\alpha_i}\left(f_{i+1} - f_i\right)^T \tilde{f}_{i+1} + \beta_i p_i^T \tilde{q}_i = -\frac{p_i^T \tilde{q}_i}{\tilde{f}_i^T f_i}\tilde{f}_{i+1}^T f_{i+1} + \frac{\tilde{f}_{i+1}^T f_{i+1}}{\tilde{f}_i^T f_i} p_i^T \tilde{q}_i = 0,
\end{aligned}$$

so that $\tilde{p}_j^T q_{i+1} = 0$ and $p_j^T \tilde{q}_{i+1} = 0$, $\forall 1 \le j \le 1$.

Remark 69 The BCG method can terminate sooner than n steps. Either $f_k = 0$ for some index $k \le n$ (so that we obtain a solution of the system $As + f = 0$ after less than n steps) or $f_k \ne 0$ and $\tilde{f}_k^T f_k = 0$ (essential breakwown common for all methods based on nonsymetric Lanczos process) or $f_k \ne 0$ and $\tilde{p}_k^T q_k = 0$ (breakdown proper of the BCG method). In usual situation breakdowns do not occur (only exceptionally). However, difficulties connected with stability can occur when $f_k \ne 0$ and $\tilde{f}_k^T f_k \approx 0$ or $f_k \ne 0$ and $\tilde{p}_k^T q_k \approx 0$.

Lemma 22 Let the assumptions of Theorem 58 be satisfied. Then the vectors f_j, $1 \le j \le i \le n$, (and also the vectors p_j, $1 \le j \le i \le n$) form a basis in the Krylov subspace

$$\mathcal{K}_i = \text{span}\{f,\ Af, \dots, A^{i-1}f\}$$

and the vectors $\tilde{f}_j$, $1 \le j \le i \le n$ (and also the vectors $\tilde{p}_j$, $1 \le j \le i \le n$) form a basis in the Krylov subspace

$$\tilde{\mathcal{K}}_i = \text{span}\{\tilde{f},\ (A^T)\tilde{f}, \dots, (A^T)^{i-1}\tilde{f}\}.$$

Proof (by induction). For $i = 1$ the statement obviously holds. Suppose that the statement holds for an index $i < n$. Since $f_i \in \mathcal{K}_i$ and $p_i \in \mathcal{K}_i$, we obtain $f_{i+1} = f_i + \alpha_i A p_i \in \mathcal{K}_{i+1}$ and $p_{i+1} = -f_{i+1} + \beta_i p_i \in \mathcal{K}_{i+1}$, and since the vectors f_j $1 \le j \le i+1$ (and also the vectors p_j, $1 \le j \le i+1$) are linearly independent (see the proof of Theorem 58), they form a basis in $\mathcal{K}_{i+1}$. Since $\tilde{f}_i \in \tilde{\mathcal{K}}_i$ and $\tilde{p}_i \in \tilde{\mathcal{K}}_i$, we obtain $\tilde{f}_{i+1} = \tilde{f}_i + \alpha_i A^T \tilde{p}_i \in \tilde{\mathcal{K}}_{i+1}$ and $\tilde{p}_{i+1} = -\tilde{f}_{i+1} + \beta_i \tilde{p}_i \in \tilde{\mathcal{K}}_{i+1}$, and since the vectors $\tilde{f}_j$, $1 \le j \le i+1$ (and also the vectors $\tilde{p}_j$, $1 \le j \le i+1$) are linearly independent (see the proof of Theorem 58), they form a basis $\tilde{\mathcal{K}}_{i+1}$.

Remark 70 Let the assumptions of Theorem 58 are satisfied. Then

$$\begin{array}{ll} f_i = \varphi_i(A)f \quad , & \tilde{f}_i = \varphi_i^T(A)\tilde{f}, \\ p_i = -\psi_i(A)f \quad , & \tilde{p}_i = -\psi_i^T(A)\tilde{f}, \end{array}$$

$\forall 1 \leq i \leq n+1$, where φ_i and ψ_i are matrix polynomials of order at most $i-1$. These polynomials can be computed using recurrence relations $\varphi_1 = I$, $\psi_1 = I$ and

$$\begin{aligned} \varphi_{i+1} &= \varphi_i - \alpha_i A \psi_i, \\ \psi_{i+1} &= \varphi_{i+1} + \beta_i \psi_i, \end{aligned}$$

$1 \leq i \leq n$. It follows immediately from the recurrence relations of the BCG method.

The coefficients α_i and β_i, $1 \leq i \leq n$, can be expressed using the polynomials φ_i and ψ_i, $1 \leq i \leq n$, so that

$$\alpha_i = \frac{\tilde{f}_i^T f_i}{\tilde{p}_i^T A p_i} = \frac{\tilde{f}^T \varphi_i^2(A) f}{\tilde{f}^T A \psi_i^2(A) f},$$

(since the matrix A and the polynomial $\psi_i(A)$ commute). Since the coefficients α_i and β_i, $1 \leq i \leq n$ can be also used for the determination of the polynomial $\varphi_i^2(A)$ and $\psi_i^2(A)$, $1 \leq i \leq n$, we can define a new iterative process $\overline{s}_i \in R^n$, $1 \leq i \leq n+1$ in such a way that $\overline{f}_i = A\overline{s}_i + f = \varphi_i^2(A)f$, $1 \leq i \leq n+1$ holds.

Lemma 23 Let the matrix polynomials φ_i and ψ_i satisfy the recurrence relations

$$\varphi_1 = I, \qquad \psi_1 = I$$

and

$$\begin{aligned} \varphi_{i+1} &= \varphi_i - \alpha_i A \psi_i, \\ \psi_{i+1} &= \varphi_{i+1} + \beta_i \psi_i \end{aligned}$$

for $1 \leq i \leq n$. Then the matrix polynomials φ_i^2 and ψ_i^2 satisfy the recurrence relations

$$\varphi_1^2 = I, \qquad \psi_1^2 = I, \qquad \varphi_1\psi_1 = I$$

and

$$\begin{aligned} \varphi_{i+1}\psi_i &= \varphi_i\psi_i - \alpha_i A \psi_i^2, \\ \varphi_{i+1}^2 &= \varphi_i^2 - \alpha_i A(\varphi_i\psi_i + \varphi_{i+1}\psi_i), \\ \varphi_{i+1}\psi_{i+1} &= \varphi_{i+1}^2 + \beta_i \varphi_{i+1}\psi_i, \\ \psi_{i+1}^2 &= \varphi_{i+1}\psi_{i+1} + \beta_i(\varphi_{i+1}\psi_i + \beta_i\psi_i^2) \end{aligned}$$

for $1 \leq i \leq n$.

Proof Multiplying the recurrence relation for φ_{i+1} by the polynomial ψ_i, we obtain

$$\varphi_{i+1}\psi_i = \varphi_i\psi_i - \alpha_i A \psi_i^2.$$

Squaring the relation for φ_{i+1}, we can write

$$\begin{aligned} \varphi_{i+1}^2 &= \varphi_i^2 - 2\alpha_i A \varphi_i \psi_i + \alpha_i^2 A^2 \psi_i^2 = \varphi_i^2 - \alpha_i A(2\varphi_i\psi_i - \alpha_i A \psi_i^2)) = \\ &= \varphi_i^2 - \alpha_i A(\varphi_i\psi_i + \varphi_{i+1}\psi_i). \end{aligned}$$

Multiplying the recurrence relation for ψ_{i+1} by the polynomial φ_{i+1}, we obtain

$$\varphi_{i+1}\psi_{i+1} = \varphi_{i+1}^2 + \beta_i\varphi_{i+1}\psi_i.$$

Squaring the relation for ψ_{i+1}, we can write

$$\begin{aligned}\psi_{i+1}^2 &= \varphi_{i+1}^2 + 2\beta_i\varphi_{i+1}\psi_i + \beta_i^2\psi_i^2 = \varphi_{i+1}^2 + \beta_i\varphi_{i+1}\psi_i + \beta_i(\varphi_{i+1}\psi_i + \beta_i\psi_i^2) = \\ &= \varphi_{i+1}\psi_{i+1} + \beta_i(\varphi_{i+1}\psi_i + \beta_i\psi_i^2).\end{aligned}$$

If we let $\overline{f}_i = \varphi_i^2 f$, $p_i = \psi_i^2 f$, $v_i = A\psi_i^2 f = Ap_i$, $u_i = \varphi_i\psi_i f$, $q_i = \varphi_{i+1}\psi_i f = u_i - \alpha_i v_i$, we get the recurrence relations, which are a basis for the CGS method.

Definition 36 Let $A \in R^{n\times n}$ be a nonsingular matrix and $f \in R^n$, $\tilde{f} \in R^n$. Then the iterative process described by the recurrence relations

$$\overline{s}_1 = 0, \qquad \overline{f}_1 = f, \qquad p_1 = f, \qquad u_1 = f$$

and

$$\begin{gathered} v_i = Ap_i, \qquad \alpha_i = \tilde{f}^T\overline{f}_i/\tilde{f}^T v_i, \\ q_i = u_i - \alpha_i v_i, \\ \overline{s}_{i+1} = \overline{s}_i - \alpha_i(u_i + q_i), \\ \overline{f}_{i+1} = \overline{f}_i - \alpha_i A(u_i + q_i), \qquad \beta_i = \tilde{f}^T\overline{f}_{i+1}/\tilde{f}^T\overline{f}_i, \\ u_{i+1} = \overline{f}_{i+1} + \beta_i q_i, \\ p_{i+1} = u_{i+1} + \beta_i(q_i + \beta_i p_i) \end{gathered}$$

for $1 \le i \le n$, is the conjugate method squared (CGS) defined by the matrix $A \in R^{n\times n}$ and the vectors $f \in R^n$, $\tilde{f} \in R^n$.

Remark 71 If the assumptions of Theorem 58 are satisfied then

$$\| \overline{f}_i \| = \| \varphi_i^2(A) f \| \le \| \varphi_i(A) \| \| \varphi_i(A) f \| = \| \varphi_i(A) \| \| f_i \|,$$

$1 \le i \le n+1$, so that the CGS method finds a solution of the system $As + f = 0$ after at most n steps ($\| f_{n+1} \| = 0$ by Theorem 58).

An advantage of the CGS methods is the fact that it does not use the transposed matrix, which is necessary for the construction of the difference version of the discrete Newton method, when the multiplication $J(x)v$ is replaced by computation of the difference $(f(x+\delta v) - f(x))/\delta$. A disadvantage of the CGS method (similarly as the BCG method) is the fact that it is not based on a suitable minimization principle. The norms of residual vectors do not monotonily decrease, they can rather oscillate. Therefore additional modifications of the CGS method, based on a residual smoothing technique, are used

Lemma 24 Let $\overline{f}_i$, $i \in N$, be a sequence of residual vectors determined by the CGS method. Let $f_1 = \overline{f}_1$ and

$$\begin{aligned}\lambda_i &= -\frac{\overline{f}_{i+1}^T(f_i - \overline{f}_{i+1})}{\| f_i - \overline{f}_{i+1} \|^2}, \\ f_{i+1} &= \overline{f}_{i+1} + \lambda_i(f_i - \overline{f}_{i+1}),\end{aligned}$$

$1 \le i \le n$. Then

$$\lambda_i = \arg\min_{\lambda\in R} \| \overline{f}_{i+1} + \lambda(f_i - \overline{f}_{i+1}) \|,$$

$1 \le i \le n$, so that $\| f_{i+1} \| \le \| f_i \|$ (norms of residual vectors are monotonely decreasing) and $\| f_{i+1} \| \le \| \overline{f}_{i+1} \|$ (the solution is found after at most n steps).

Proof For $1 \leq i \leq n$, we can write

$$\| f_{i+1} \|^2 = \| \overline{f}_{i+1} \|^2 + 2\lambda_i \overline{f}_{i+1}^T (f_i - \overline{f}_{i+1}) + \lambda_i^2 \| f_i - \overline{f}_{i+1} \|^2 .$$

This quadratic function reaches its minimun for $\lambda_i = -\overline{f}_{i+1}^T (f_i - \overline{f}_{i+1}) / \| f_i - \overline{f}_{i+1} \|^2$.

Recurrence relations for f_i (Lemma 24) together with corresponding recurrence relations for s_i are a basis for the single smoothed CGS method.

Definition 37 Let $A \in R^{n \times n}$ be a nonsingular matrix and $f \in R^n$, $\tilde{f} \in R^n$. Then the iterative process described by the recurrence relations

$$s_1 = 0, \quad \overline{s}_1 = 0, \quad f_1 = f, \quad \overline{f}_1 = f, \quad p_1 = f, \quad u_1 = f$$

and

$$\begin{gathered}
v_i = Ap_i, \quad \alpha_i = \tilde{f}^T f_i / \tilde{f}^T v_i, \\
q_i = u_i - \alpha_i v_i, \\
\overline{s}_{i+1} = \overline{s}_i - \alpha_i (u_i + q_i), \\
\overline{f}_{i+1} = \overline{f}_i - \alpha_i A (u_i + q_i), \quad \beta_i = \tilde{f}^T f_{i+1} / \tilde{f}^T \overline{f}_i, \\
u_{i+1} = \overline{f}_{i+1} + \beta_i q_i, \\
p_{i+1} = u_{i+1} + \beta_i (q_i + \beta_i p_i), \\
\lambda_i = -\frac{\overline{f}_{i+1}^T (f_i - \overline{f}_{i+1})}{\| f_i - \overline{f}_{i+1} \|^2}, \\
s_{i+1} = \overline{s}_{i+1} + \lambda_i (s_i - \overline{s}_{i+1}), \\
f_{i+1} = \overline{f}_{i+1} + \lambda_i (f_i - \overline{f}_{i+1}),
\end{gathered}$$

for $1 \leq i \leq n$, is the single smoothed CGS (SSCGS) method defined by the matrix $A \in R^{n \times n}$ and the vectors $f \in R^n$, $\tilde{f} \in R^n$.

Although the norms of the residual vectors of the single smoothed CGS method are monotone decreasing, the doubly smoothed CGS method is more advantageous for construction of trust region methods.

Definition 38 Let $A \in R^{n \times n}$ be a nonsingular matrix and $f \in R^n$, $\tilde{f} \in R^n$. Then the iterative proces described by the recurrence relations

$$s_1 = 0, \quad \overline{s}_1 = 0, \quad f_1 = f, \quad \overline{f}_1 = f, \quad p_1 = f, \quad u_1 = f$$

and

$$\begin{gathered}
v_i = Ap_i, \quad \alpha_i = \tilde{f}^T \overline{f}_i / \overline{f}^T v_i, \\
q_i = u_i - \alpha_i v_i, \\
\overline{s}_{i+1} = \overline{s}_i - \alpha_i (u_i + q_i), \\
\overline{f}_{i+1} = \overline{f}_i - \alpha_i A (u_i + q_i), \quad \beta_i = \tilde{f}^T \overline{f}_{i+1} / \tilde{f}^T \overline{f}_i, \\
u_{i+1} = \overline{f}_{i+1} + \beta_i q_i, \\
p_{i+1} = u_{i+1} + \beta_i (q_i + \beta_i p_i), \\
[\lambda_i, \mu_i]^T = \arg \min_{[\lambda, \mu]^T \in R^2} \| \overline{f}_{i+1} + \lambda (f_i - \overline{f}_{i+1}) + \mu v_i \|, \\
s_{i+1} = \overline{s}_{i+1} + \lambda_i (s_i - \overline{s}_{i+1}) + \mu_i p_i, \\
f_{i+1} = \overline{f}_{i+1} + \lambda_i (f_i - \overline{f}_{i+1}) + \mu_i v_i,
\end{gathered}$$

for $1 \leq i \leq n$, is the doubly smoothed CGS (DSCGS) method defined by the matrix $A \in R^{n \times n}$ and the vectors $f \in R^n$, $\tilde{f} \in R^n$.

Remark 72 The vector $[\lambda_i, \mu_i]^T$ realizing a minimum of the norm $\| f_{i+1} \|$ can be determined by the formula

$$\begin{bmatrix} \lambda_i \\ \mu_i \end{bmatrix} = -(V_i^T V_i)^{-1} V_i^T \overline{f}_{i+1},$$

where $V_i = [f_i - \overline{f}_{i+1}, v_i] \in R^{n \times 2}$ (derivation of this formula is analogous to the derivation of the formula for λ_i in Lemma 24). Substituting this expression into the relation for f_{i+1}, we obtain $f_{i+1} = P_i \overline{f}_{i+1}$, where $P_i = I - V_i(V_i^T V_i)^{-1} V_i^T$ is the matrix of orthogonal projection into the subspace spanned by vectors $f_i - \overline{f}_{i+1}$ and v_i.

The CGS, SSCGS, DSCGS methods can be modified using preconditioning. Since, for inexact solution of the system $As + f = 0$, we require the residual vector $As + f$, we have to use the right preconditioning. This means that the system $AC^{-1}\hat{s} + f = 0$ with a preconditioner C^{-1} is solved and then $s = C^{-1}\hat{s}$ is set. Since modifications of all the CGS, the SSCGS and the DSCGS methods are practically the same, we introduce only preconditioned version of the DSCGS method, which is described by the recurrence relations

$$s_1 = 0, \quad \overline{s}_1 = 0, \quad f_1 = f, \quad \overline{f}_1 = f, \quad p_1 = f, \quad u_1 = f$$

and

$$\begin{aligned} v_i = AC^{-1}p_i, &\quad \alpha_i = \tilde{f}^T \overline{f}_i / \tilde{f}^T v_i, \\ q_i &= u_i - \alpha_i v_i, \\ \overline{s}_{i+1} &= \overline{s}_i + \alpha_i C^{-1}(u_i + q_i), \\ \overline{f}_{i+1} = \overline{f}_i + \alpha_i AC^{-1}(u_i + q_i), &\quad \beta_i = \tilde{f}^T \overline{f}_{i+1} / \tilde{f}^T \overline{f}_i, \\ u_{i+1} &= \overline{f}_{i+1} + \beta_i q_i, \\ p_{i+1} &= u_{i+1} + \beta_i(q_i + \beta_i p_i), \\ [\lambda_i, \mu_i]^T &= -(V_i^T V_i)^{-1} V_i^T \overline{f}_{i+1}, \\ s_{i+1} &= \overline{s}_{i+1} + \lambda_i(s_i - \overline{s}_{i+1}) + \mu_i C^{-1} p_i, \\ f_{i+1} &= \overline{f}_{i+1} + \lambda_i(f_i - \overline{f}_{i+1}) + \mu_i v_i, \end{aligned}$$

for $1 \le i \le n$, (here $V_i = [f_i - \overline{f}_{i+1}, v_i] \in R^{n \times 2}$).

The preconditioner is usually chosen in such a way that $C \approx A$. Then the matrix $AC^{-1} \approx I$ is well-conditioned. Very efficient preconditioners are obtained using an incomplete LU decomposition

$$P(A + E) = LU$$

where L is a lower triangular matrix with units on the diagonal, U is an upper triangular matrix, P is a permutation matrix and E is a matrix resulting by a cancellation of the new nonzero elements (fill-in). The permutation matrix is chosen in such a way that the matrix PA has nonzero elements (pivots) on the diagonal.

Now we consider the GMRES method, which uses long recurrences. The idea of the GMRES method consists in the generation of orthonormal vectors q_i $1 \le i \le n$, such that q_j, $1 \le j \le i$, form an orthogonal basis in the Krylov subspace $\mathcal{K}_i$. The vector $s_{i+1} \in R^n$ is chosen so that,

$$s_{i+1} = \arg \min_{s \in \mathcal{K}_i} \| As + f \| \tag{M}$$

holds. Thus the GMRES method is based on the minimization principle, which means that norms of the residual vectors are monotonically decreasing.

The orthonormal vectors q_i $1 \le j \le n$ are generated using the Gramm-Schmidt ortogonalization process. The classic Gramm-Schmidt ortogonalization process utilizes the recurrence relations

$$\beta_1 q_1 = f$$

and

$$\begin{gathered} q_{i+1}^1 = Aq_i, \\ \left.\begin{aligned} \alpha_{ji} &= q_j^T q_{i+1}^1, \\ q_{i+1}^{j+1} &= q_{i+1}^j - \alpha_{ji} q_j, \end{aligned}\right\} 1 \le j \le i \\ \beta_{i+1} q_{i+1} = q_{i+1}^{i+1}, \end{gathered}$$

$1 \le i \le n-1$, where the coefficients β_i, $1 \le i \le n$ are chosen in such a way that the vectors q_i, $1 \le i \le n$, have unity norms. More stable is the modified Gramm-Schmidt ortogonalization process

$$\beta_1 q_1 = f$$

and

$$\begin{gathered} q_{i+1}^1 = Aq_i, \\ \left.\begin{aligned} \alpha_{ji} &= q_j^T q_{i+1}^j, \\ q_{i+1}^{j+1} &= q_{i+1}^j - \alpha_{ji} q_j, \end{aligned}\right\} 1 \le j \le i \\ \beta_{i+1} q_{i+1} = q_{i+1}^{i+1}, \end{gathered}$$

$1 \le i \le n-1$. The Gramm-Schmidt ortogonalization process generating orthonormal bases of the Krylov subspaces $\mathcal{K}_i$, $1 \le i \le n$, is usually called the Arnoldi process defined by the matrix $A \in R^{n \times n}$ and the vector $f \in R^n$.

Denoting $Q_i = [q_1, q_2, \ldots, q_i]$ and

$$H_i = \begin{bmatrix} \alpha_{11}, & \alpha_{12}, & \ldots, & \alpha_{1i} \\ \beta_2, & \alpha_{22}, & \ldots, & \alpha_{2i} \\ 0, & \beta_3, & \ldots, & \alpha_{3i} \\ \ldots & \ldots & \ldots & \ldots \\ 0, & 0, & \ldots, & \beta_{i+1} \end{bmatrix}$$

($H_i \in R^{(i+1) \times i}$ is an upper Hessenberg matrix), we can write the Arnoldi process in the matrix form

$$AQ_i = Q_{i+1} H_i.$$

Setting $s_{i+1} = Q_i z_i$, where $z_i \in R^n$, we obtain

$$\| As_{i+1} + f \| = \| AQ_i z_i + f \| = \| Q_{i+1} H_i z_i + Q_{i+1}(\beta_1 e_1) \| = \| H_i z_i + \beta_1 e_1 \|,$$

so that the minimization condition can be written in the form

$$z_i = \arg \min_{z \in R^i} \| H_i z + \beta_1 e_1 \| \qquad (\overline{\mathrm{M}}).$$

Theorem 59 Let $\mathcal{K}_j \ne \mathcal{K}_{j+1}$, $1 \le j \le i$, $\mathcal{K}_i = \mathcal{K}_{i+1}$ and let (M) hold. Then $As_{i+1} + f = 0$.

Proof Consider the Arnoldi process defined by the nonsingular matrix $A \in R^{n \times n}$ and the vector f. If $\mathcal{K}_j \ne \mathcal{K}_{j+1}$, $1 \le j \le i$ and $\mathcal{K}_i = \mathcal{K}_{i+1}$, then the vectors q_i, $1 \le j \le i$, are linearly independent and $\beta_{i+1} = 0$. Therefore, it holds

$$AQ_i = Q_i \overline{H}_i$$

where $\overline{H}_i \in R^{i \times i}$ is an upper Hessenberg matrix, which is obtained from the matrix $H_i \in R^{(i+1) \times i}$ by deleting its last row. Since the matrix AQ_i has linearly independent columns and A is nonsingular, the matrix $\overline{H}_i$ is nonsingular and a solution of the system $\overline{H}_i z_i + \beta_1 e_1 = 0$ exists. Setting $s_{i+1} = Q_i z_i$ we obtain

$$\| As_{i+1} + f \| = \| \overline{H}_i z_i + \beta_1 e_1 \| = 0$$

Corollary The GMRES method finds a solution of the system $As + f = 0$ after at most n steps. Indeed, if $\mathcal{K}_j \neq \mathcal{K}_{j+1}$, $1 \leq j < n$, then necessarily $\mathcal{K}_n = \mathcal{K}_{n+1} = R^n$. The GMRES methods cannot fail, since $\beta_{i+1} = 0$ implies $As_{i+1} + f = 0$.

For determining the vector z_i satisfying the condition ($\overline{\mathrm{M}}$), we require to compute the orthogonal decomposition

$$P_i(H_i z_i + \beta_1 e_1) = \begin{bmatrix} R_i \\ 0 \end{bmatrix} z_i + \begin{bmatrix} h_i \\ \overline{\eta}_{i+1} \end{bmatrix}$$

where $P_i = \overline{P}_i \overline{P}_{i-1} \dots \overline{P}_1$ is a product of Givens elementary rotations and

$$R_i = \begin{bmatrix} \rho_{11}, & \rho_{12}, & \dots, & \rho_{1i} \\ 0, & \rho_{22}, & \dots, & \rho_{2i} \\ 0, & 0, & \dots, & \rho_{ii} \end{bmatrix}, \qquad h_i = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \dots \\ \eta_i \end{bmatrix}.$$

This is a way which was already used in the LSQR method (Section 4.7). Therefore, we do not repeat the derivation. However, we introduce the resulting recurrence relations for the GMRES method.

Definition 39 Let $A \in R^{n \times n}$ be a nonsingular matrix and $f \in R^n$. Then the iterative process described by the recurrence relations

$$\beta_1 q_1 = f, \qquad \overline{\eta}_1 = \beta_1$$

and

$$q_{i+1}^1 = Aq_i,$$

$$\overline{\alpha}_{1i} = q_1^T q_{i+1}^1, \qquad q_{i+1}^2 = q_{i+1}^1 - \overline{\alpha}_{1i} q_1,$$

$$\left.\begin{array}{r} \alpha_{ji} = q_j^T q_{i+1}^j, \qquad q_{i+1}^{j+1} = q_{i+1}^j - \alpha_{ji} q_j, \\ \rho_{j-1i} = \lambda_{j-1}\overline{\alpha}_{j-1i} + \tau_{j-1}\alpha_{ji}, \\ \overline{\alpha}_{ji} = -\lambda_{j-1}\overline{\alpha}_{j-1i} + \tau_{j-1}\alpha_{ji}, \end{array}\right\} 1 < j \leq i$$

$$\beta_{i+1} q_{i+1} = q_{i+1}^{i+1},$$

$$\rho_{ii} = \sqrt{\overline{\alpha}_{ii}^2 + \beta_{i+1}^2},$$

$$\lambda_i = \frac{\overline{\alpha}_{ii}}{\rho_{ii}}, \qquad \tau_i = \frac{\beta_{i+1}}{\rho_{ii}},$$

$$\eta_i = \lambda_i \overline{\eta}_i, \qquad \overline{\eta}_{i+1} = -\tau_i \overline{\eta}_i,$$

$1 \leq i \leq n$, is the GMRES method defined by the matrix $A \in R^{n \times n}$ and the vector $f \in R^n$.

If we use the GMRES method, then we can write the minimization condition in the form

$$z_i = \arg\min_{z \in R^n} \left\| \begin{bmatrix} R_i \\ 0 \end{bmatrix} z + \begin{bmatrix} h_i \\ \overline{\eta}_{i+1} \end{bmatrix} \right\|.$$

Therefore $R_i z_i + h_i = 0$ holds (the matrix R_i is upper triangular) and setting $s_{i+1} = Q_i z_i$, we obtain $\| As_{i+1} + f \| = \overline{\eta}_{i+1}$. Therefore, the numbers $\overline{\eta}_i$, $1 \leq i \leq n+1$, are norms of the residual vectors $f_i = As_i + f$, $1 \leq i \leq n+1$. If the GMRES methods reaches a sufficienly small residual norm ($\overline{\eta}_{i+1} \leq \varpi \| f \|$), then we can terminate the iterative process and set $s_{i+1} = Q_i z_i$, where $R_i z_i + h_i = 0$.

The GMRES methods can be modified by various ways. If we generate orthonormal bases in the shifted Krylov subspaces

$$A\mathcal{K}_i = \operatorname{span}\{Af, \dots, A^i f\},$$

then the use of the orthogonal decomposition is not required. The vectors q_j, $1 \leq j \leq i$ are again determined by the Gramm-Schmid ortogonalization process, so that

$$AQ_{i-1} = Q_i H_{i-1},$$

where $H_{i-1} \in R^{i\times(i-1)}$ is an upper Hessenberg matrix. Choosing the vector q_1 such that $\beta_1 q_1 = Af$, we can write

$$[Af, AQ_{i-1}] = Q_i[\beta_1 e_1, H_{i-1}] = Q_i R_i,$$

where

$$R_i = \begin{bmatrix} \beta_1, & \alpha_{11}, & \alpha_{12}, & \dots, & \alpha_{1i-1} \\ 0, & \beta_2, & \alpha_{22}, & \dots, & \alpha_{2i-1} \\ 0, & 0, & \beta_3, & \dots, & \alpha_{3i-1} \\ \dots & \dots & \dots & \dots & \dots \\ 0, & 0, & 0, & \dots, & \beta_i \end{bmatrix}$$

($R_i \in R^{i\times i}$ is an upper triangular matrix). Setting

$$s_{i+1} = [f, Q_{i-1}]z_i,$$

we obtain $s_{i+1} \in \mathcal{K}_i$, since the vectors f and q_j, $1 \le j \le i-1$, are linearly independent. Furthermore, it holds

$$\| As_{i+1} + f \| = \| [Af, AQ_{i-1}]z_i + f \| = \| Q_i R_i z_i + f \|,$$

so that the minimization condition can be written in the form

$$z_i = \arg\min_{z\in R^i} \| Q_i R_i z + f \| .$$

The system of normal equations for this least squares problem has the form $R_i^T Q_i^T Q_i R_i z_i + R_i^T Q_i^T f = 0$, so that

$$R_i z_i + Q_i^T f = 0,$$

which, after substitution into the formula for the residual vector, gives

$$f_{i+1} = As_{i+1} + f = (I - Q_i Q_i^T)f = f_i - q_i q_i^T f.$$

Since the orthogonality implies $q_i^T Q_{i-1} = 0$, we can write $q_i^T f_i = q_i^T(I - Q_{i-1}Q_{i-1}^T)f = q_i^T f$, which gives

$$f_{i+1} = f_i - q_i q_i^T f_i.$$

This formula improves the stability of the modified GMRES method. Utilizing all above results, we can give the definition of the modified GMRES method.

Definition 40 Let $A \in R^{n\times n}$ be a nonsingular matrix and $f \in R^n$. Then the iterative process described by the recurrence relations

$$f_1 = f, \qquad \beta_1 q_1 = Af$$

and

$$\begin{gathered} \gamma_i = q_i^T f_i, \\ f_{i+1} = f_i - \gamma_i q_i, \\ q_{i+1}^1 = Aq_i, \\ \left.\begin{aligned} \alpha_{ji} &= q_j^T q_{i+1}^j, \\ q_{i+1}^{j+1} &= q_{i+1}^j - \alpha_{ji} q_j, \end{aligned}\right\} 1 \le j \le i \\ \beta_{i+1} q_{i+1} = q_{i+1}^{i+1}, \end{gathered}$$

$1 \le i \le n-1$, is the modified GMRES method defined by the matrix $A \in R^{n\times n}$ and the vector $f \in R^n$.

If the modified GMRES method reaches a sufficiently small residual norm ($\| f_{i+1} \| \leq \overline{\omega} \| f \|$), we can terminate the iterative process and set $s_{i+1} = [f, Q_{i-1}] z_i$, where

$$\begin{bmatrix} \beta_1, & \alpha_{11}, & \dots, & \alpha_{1i-1} \\ 0, & \beta_2, & \dots, & \alpha_{2i-1} \\ \dots & \dots & \dots & \dots \\ 0, & 0, & \dots, & \beta_i \end{bmatrix} z_i = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \dots \\ \gamma_i \end{bmatrix}.$$

Remark 73 The basic and the modified GMRES methods can be easily preconditioned (the right preconditioning is used). In this case, the matrix AC^{-1} is used instead of the matrix A and the vector $s_{i+1} \in R^n$ is determined by the formula

$$s_{i+1} = -C^{-1} Q_i R_i^{-1} h_i$$

(the basic method) nebo

$$s_{i+1} = -C^{-1} [f, Q_{i-1}] R_i^{-1} Q_i^T f$$

(the modified method). The preconditioner C^{-1} is again chosen to satisfy $C \approx A$.

6.8 Inexact trust region methods

Remark 74 Both the smoothed CGS and the GMRES methods can be used for the construction of the inexact trust region method. In this case, a sequence of vectors $s_{i+1} \in R^n$, $1 \leq i \leq n$, approximating the solution of the system $As + f = 0$, is generated and either $s = s_{i+1}$ is set, if $\| s_{i+1} \| \leq \Delta$ and $\| f_{i+1} \| \leq \overline{\omega} \| f \|$, $0 \leq \overline{\omega} < 1$, or $s = s_i + \alpha_i (s_{i+1} - s_i)$ with $\| s \| = \Delta$ is set, if $\| s_j \| < \Delta$, $\| f_j \| > \overline{\omega} \| f \|$, $1 \leq j \leq i$, and $\| s_{i+1} \| \geq \Delta$. This choice obviously satisfies the conditions (T1a), (T1b) required by a trust region method (Definition 33). We require to formulate the assumptions guaranteeing the condition (T1c) to hold or

$$\| f \| - \| As + f \| \geq 2\underline{\sigma} \| As \|),$$

where $\underline{\sigma}$ is some constant. In the sequel, we suppose that the matrix A satisfy the condition $\|I - A\| \leq \overline{\nu} < 1$, which can be obtained by a suitable preconditioning (the matrix A is replaced by the matrix AC^{-1} such that $\|I - AC^{-1}\| \leq \overline{\nu} < 1$).

Lemma 25 Let $\|I - A\| \leq \overline{\nu} < 1$ and let $s_{i+1} \in R^n$, $i = 1, \dots, n$, be the vectors generated by the GMRES method or by the doubly smoothed CGS method. Then

$$\|f\|^2 - \|r_{i+1}\|^2 \geq \underline{\eta}^2 \|f\|^2,$$

where $\underline{\eta} = (1 - \overline{\nu})/(1 + \overline{\nu})$.

Proof (a) First we prove that

$$|f^T A f| \geq \frac{1 - \overline{\nu}}{1 + \overline{\nu}} \|f\| \|Af\| = \underline{\eta} \|f\| \|Af\|.$$

From the assumption it follows

$$\begin{aligned} |f^T A f| &= |f^T f - f^T (I - A) f| \geq |f^T f| - |f^T (I - A) f| \\ &\geq \|f\|^2 - \|I - A\| \|f\|^2 \geq (1 - \overline{\nu}) \|f\|^2 \end{aligned}$$

and

$$\|Af\| \leq \|f\| + \|I - A\| \|f\| \leq (1 + \overline{\nu}) \|f\|,$$

which together give the required inequality.
(b) Since the sequence of norms of the residual vectors generated by the GMRES method or by the doubly smoothed CGS method is nonincreasing, it suffices to prove that

$$\|f\|^2 - \|r_2\|^2 \geq \underline{\eta}^2 \|f\|^2.$$

Consider the GMRES method first. Since $s_1 = 0$ and $\mathcal{K}_1 = \mathrm{span}\{f\}$, it holds

$$\|r_2\| = \min_{\mu \in R} \|A(\mu f) + f\|.$$

From the optimality condition

$$\mu_1 \stackrel{\Delta}{=} \arg\min_{\mu \in R} \|A(\mu f) + f\|^2 = \arg\min_{\mu \in R} (\mu^2 \|Af\|^2 + 2\mu f^T A f + \|f\|^2),$$

we obtain $\mu_1 = -f^T Af/\|Af\|^2$ so that the residual norm r_2 satisfies the equality

$$\|r_2\|^2 = \frac{(f^T Af)^2}{\|Af\|^4}\|Af\|^2 - 2\frac{(f^T Af)^2}{\|Af\|^2} + \|f\|^2 = \|f\|^2 - \frac{(f^T Af)^2}{\|Af\|^2\|f\|^2}\|f\|^2.$$

This equality together with (a) proves the statement of the lemma for the GMRES method. Consider now the doubly smoothed CGS method. Then

$$\|r_2\| = \min_{[\lambda,\mu]^T \in R^2} \|\overline{r}_2 + \lambda(f - \overline{r}_2) + \mu v_1\| \le \min_{\mu \in R} \|f + \mu v_1\| = \min_{\mu \in R} \|f + \mu Af\|$$

(after substituting $\lambda = 1$), which gives the same result as in the case of the GMRES method.

Lemma 26 Let the assumptions of Lemma 25 be satisfied and let $s \in \mathcal{R}^n$ be the vector determined by the GMRES method or by the doubly smoothed CGS method as in Remark 74. Then

$$\|f\| - \|As + f\| \ge 2\underline{\sigma}\|As\|,$$

where $2\underline{\sigma} = \underline{\eta}^2/8$.

Proof (a) Let $\|s_{i+1}\| < \Delta$ and $\|r_{i+1}\| \le \overline{\omega}\|f\|$. Then Lemma 25 implies

$$2\|f\|\left(\|f\| - \|r_{i+1}\|\right) \ge \|f\|^2 - \|r_{i+1}\|^2 \ge \underline{\eta}^2\|f\|^2,$$

which together with the inequality $\underline{A}\|s\| \le \|As\| \le 2\|f\|$ gives

$$\|f\| - \|r_{i+1}\| \ge \frac{1}{2}\underline{\eta}^2\|f\| \ge \frac{1}{4}\underline{\eta}^2\|As\|.$$

(b) Let $\|s_{i+1}\| \ge \Delta$ and $i > 1$. Then $s = \tau_i s_{i+1} + (1 - \tau_i)s_i$ with $0 < \tau_i \le 1$ holds, so that

$$\|As + f\| = \|\tau_i(As_{i+1} + f) + (1 - \tau_i)(As_i + f)\| \le \tau_i\|r_{i+1}\| + (1 - \tau_i)\|r_i\|$$

and Lemma 25 together with the inequality $\underline{A}\|s\| \le \|As\| \le 2\|f\|$ implies

$$\|f\| - \|As + f\| \ge \tau_i\left(\|f\| - \|r_{i+1}\|\right) + (1 - \tau_i\left(\|f\| - \|r_i\|\right) \ge \frac{1}{2}\underline{\eta}^2\|f\| \ge \frac{1}{4}\underline{\eta}^2\|As\|.$$

(c) Let $\|s_{i+1}\| \ge \Delta$ and $i = 1$. Then $s = \tau_1 s_2$ holds, where $0 < \tau_1 \le 1$. Therefore, we can write

$$\begin{aligned}
\|f\|^2 - \|As + f\|^2 &= \|f\|^2 - \tau_1^2\|As_2\|^2 - 2\tau_1 f^T As_2 - \|f\|^2 \\
&= -\tau_1^2\|As_2\|^2 - 2\tau_1 f^T As_2 \ge \tau_1\left(-\|As_2\|^2 - 2f^T As_2\right) \\
&= \tau_1\left(\|f\|^2 - \|As_2 + f\|^2\right)
\end{aligned}$$

(since $\tau_1^2 \le \tau_1$ pro $0 < \tau_1 \le 1$), or

$$\begin{aligned}
2\|f\|\left(\|f\| - \|As + f\|\right) &\ge \|f\|^2 - \|As + f\|^2 \ge \tau_1\left(\|f\|^2 - \|r_2\|^2\right) \\
&\ge \tau_1\|f\|\left(\|f\| - \|r_2\|\right),
\end{aligned}$$

so that

$$\|f\| - \|As + f\| \ge \frac{1}{2}\tau_1\left(\|f\| - \|r_2\|\right) \ge \frac{1}{4}\tau_1\underline{\eta}^2\|f\|$$

as in the case (a). Therefore, it holds

$$2\|f\| \ge \|r_2 - f\| = \|As_2\|,$$

which after substitution into the previous inequality gives

$$\|f\| - \|As + f\| \ge \frac{1}{8}\tau_1\underline{\eta}^2\|As_2\| = \frac{1}{8}\underline{\eta}^2\|As\|.$$

Theorem 60 Let $\|I - A_i\| \leq \overline{\nu} < 1$, $i \in N$ and let $s_i \in R^n$, $i \in N$, are direction vectors determined by the GMRES method or by the doubly smoothed CGS method as in Remark 74. Then the conditions (T1a)-(T1c) are satisfied and the direction vectors $s_i \in R^n$, $i \in N$, can be used for the construction of an inexact trust region method. If this method is applied to the mapping $f : R^n \to R^n$ satisfying the assumption (J3)-(J5) and if the matrices A_i, $i \in N$ satisfy the conditions (A3)-(A4), then $x_i \to x^\star$ and $f(x^\star) = 0$.

Proof This theorem is an immediate consequence of Lemma 26 and Theorem 48.

The GMRES method or the doubly smoothed method can be also used for the construction of dog-leg methods. In this case the vectors $s_{i+1} \in R^n$, $1 \leq i \leq m$ are generated, where $m \ll n$ (usually $1 \leq m \leq 3$). If for some index $1 \leq i \leq m$ $\| s_{i+1} \| \leq \Delta$ and $\| f_{i+1} \| \leq \overline{\omega} \| f \|$, $0 \leq \overline{\omega} < 1$ hold, then we set $s = s_{i+1}$. If for some index $1 \leq i \leq m$ $\| s_j \| < \Delta$, $\| f_j \| > \overline{\omega} \| f \|$, $1 \leq j \leq i$, and $\| s_{i+1} \| \geq \Delta$ hold, then we set $s = s_i + \alpha_i(s_{i+1} - s_i)$, so that $\| s \| = \Delta$. If none of these cases is true, then we determine the solution $s^* \in R^n$ of the system $As + f = 0$ using some direct elimination method and set $s = s_{m+1} + \alpha_{m+1}(s^* - s_{m+1})$. It is easy to prove (similarly to the proof of Lemma 26) that if $\Delta \geq \underline{\gamma} \| f \|$ or $|f^T A f| \geq \underline{\varepsilon} \| f \| \| A f \|$ holds, then the condition $(\overline{\text{T1c}})$ is satisfied.

The following table shows a comparison of several realization of the sparse discrete Newton method (DSCGS denotes the doubly smoothed CGS method, GMRES(30) or GMRES(10) denotes the GMRES method restarted always after 30 or 10 iterations of the Arnoldi process, L denotes the line search method, T denotes the trust region method and LU denotes the preconditioning by the incomplete LU decomposition) for solving 18 sparse systems of nonlinear equations with 100 unknowns (the total number of iterations NIT and function evaluations NFV is given together with the total computational time).

Method	NIT-NFV	time
Newton + DSCGS (L)	330-2010	8.07
Newton + DSCGS (L + LU)	235-1184	3.79
Newton + GMRES(30) (L)	346-2081	14.78
Newton + GMRES(10) (L + LU)	235-1184	3.85
Newton direct (L + complete LU)	238-1200	5.21
Newton + DSCGS (T)	431-1851	8.96
Newton + DSCGS (T + LU)	234-1050	3.79
Newton + GMRES(30) (T)	337-1570	13.07
Newton + GMRES(10) (T + LU)	236-1061	3.95
Newton direct (T + complete LU)	221- 975	5.00

The next table shows a comparison of several methods (the sparse discrete Newton method, the Schubert quasi-Newton method, the 5-step limited storage Broyden method, the difference version of the inexact Newton method, the Li method with the Schubert update), realized as inexact trust region methods based on unpreconditioned DSCGS iterations, for solving 18 sparse systems of nonlinear equations with 100 unknowns (the total number of iterations NIT and function evaluations NFV is given together with the total computational time).

Method	NIT-NFV	time
discrete Newton	389-1784	7.58
Schubert	739-1288	10.98
5 - Broyden	647-1322	12.03
inexact Newton	517-6668	15.65
Li + Schubert	681-1592	11.09

7 Optimization of dynamical systems

Consider a problem with the objective function

$$F(x) = \int_{t_0}^{t_1} f_A(y(x,t),t)\,dt + f_T(y(x,t_1)), \tag{O}$$

where

$$\frac{dy(x,t)}{dt} = f_S(x, y(x,t), t), \qquad y(x,t_0) = f_I(x)\ . \tag{D}$$

Here $x \in R^n$, $y : R^n \times [t_0,t_1] \to R^{ns}$, $F : R^n \to R$, $f_A : R^{ns} \times [t_0,t_1] \to R$, $f_T : R^{ns} \to R$, $f_S : R^n \times R^{ns} \times [t_0,t_1] \to R^{ns}$, $f_I : R^n \to R^{ns}$. The integral (O) can be replaced by an additional differential equation. Then

$$F(x) = F_A(x,t_1) + f_T(y(x,t_1)), \tag{$\overline{\mathrm{O}}$}$$

where

$$\frac{dy(x,t)}{dt} = f_S(x,y,t), \qquad y(x,t_0) = f_I(x),$$

$$\frac{dF_A(x,t)}{dt} = f_A(y,t), \qquad F_A(x,t_0) = 0. \tag{$\overline{\mathrm{D}}$}$$

Totally, $n_S + 1$ differential equations are solved in the forward direction. It suffices to compute terminal values. The problem ($\overline{\mathrm{O}}$) + ($\overline{\mathrm{D}}$) can be solved using gradient optimization methods (CG, VM, N). For this reason, we require to compute derivatives of the objective function.

Assumptions:

(A1) A continuous solution of the system (D) on the interval $[t_0,t_1]$ exists whenever $x \in X \subset R^n$.

(A2) The functions f_A, f_T, f_S, f_I are twice continuously differentiable on $X \subset R^n$.

Here $X \subset R^n$ is a region containing all points $x_i \in R^n$ $i \in N$, reached during iterative process.

7.1 Forward computation of the gradient

Denote $u(x,t) = dy(x,t)/dx$, so that $u : R^n \times [t_0,t_1] \to R^{ns\times n}$. Differentiating ($\overline{\mathrm{O}}$) and ($\overline{\mathrm{D}}$) we obtain

$$g^T(x) = g_A^T(x,t_1) + \frac{\partial f_T(y(x,t_1))}{\partial y} u(x,t_1), \tag{$\overline{\mathrm{O1}}$}$$

where

$$\frac{du(x,t)}{dt} = \frac{\partial f_S(x,y,t)}{\partial y} u(x,t) + \frac{\partial f_S(x,y,t)}{\partial x}, \qquad u(x,t_0) = \frac{df_I(x)}{dx},$$

$$\frac{dg_A^T(x,t)}{dt} = \frac{\partial f_A(y,t)}{\partial y} u(x,t), \qquad g_A^T(x,t_0) = 0. \tag{$\overline{\mathrm{D1}}$}$$

Here $g^T(x) = dF(x)/dx$, $g_A^T(x,t) = dF_A(x,t)/dx$. Totally, $(n_S+1)(n+1)$ differential equations are solved in the forward direction.

7.2 Backward computation of the gradient

Let $p(t)$ be an arbitrary function, such that $p : [t_0, t_1] \to R^{n_S}$ and let $y(x,t)$ be a solution of system (D), so that $f_S(x,y,t) - dy(x,t)/dt = 0$ for $t \in [t_0, t_1]$. Using (O), we can write

$$F(x) = \int_{t_0}^{t_1} \{f_A(y,t) + p^T(t)(f_S(x,y,t) - \frac{dy(x,t)}{dt})\}\, dt + f_T(y(x,t_1))$$

and using per partes integration we obtain

$$\begin{aligned} F(x) \quad &= \quad \int_{t_0}^{t_1} \{f_A(y,t) + p^T(t) f_S(x,y,t) + \frac{dp^T(t)}{dt} y(x,t)\}\, dt \\ &\quad + p^T(t_0) y(x,t_0) - p^T(t_1) y(x,t_1) + f_T(y(x,t_1)) \; . \end{aligned}$$

Now the function $F(x)$ can be differentiated with respect to x, so that

$$\begin{aligned} g^T(x) \quad &= \quad \int_{t_0}^{t_1} \left\{ \left[\frac{\partial f_A(y,t)}{\partial y} + p^T(t) \frac{\partial f_S(x,y,t)}{\partial y} + \frac{dp^T(t)}{dt} \right] \frac{dy(x,t)}{dx} \right. \\ &\quad \left. + \; p^T(t) \frac{\partial f_S(x,y,t)}{\partial x} \right\} dt \\ &\quad + p^T(t_0) \frac{df_I(x)}{dx} + \left[\frac{\partial f_T(y(x,t_1))}{\partial y} - p^T(t_1) \right] \frac{dy(x,t_1)}{dx} \; . \end{aligned}$$

Choosing the function $p(t)$ in such a way that all terms with $dy(x,t)/dt$ cancel, i.e. so that

$$-\frac{dp(x,t)}{dt} = (\frac{\partial f_S(x,y,t)}{\partial y})^T p(x,t) + (\frac{\partial f_A(y,t)}{\partial y})^T, \qquad p(x,t_1) = (\frac{\partial f_T(y(x,t_1))}{\partial y})^T,$$

we can write

$$g^T(x) = \int_{t_0}^{t_1} p^T(x,t) \frac{\partial f_S(x,y,t)}{\partial x}\, dt + p^T(x,t_0) \frac{df_I(x)}{dx} \; .$$

The above results can be summarized in the form

$$g(x) = \tilde{g}_A(x,t_0) + (\frac{df_I(x)}{dx})^T p(x,t_0), \tag{$\overline{\text{O2}}$}$$

where

$$-\frac{dp(x,t)}{dt} = (\frac{\partial f_S(x,y,t)}{\partial y})^T p(x,t) + (\frac{\partial f_A(y,t)}{\partial y})^T, \qquad p(x,t_1) = (\frac{\partial f_T(y(x,t_1))}{\partial y})^T$$

and

$$-\frac{d\tilde{g}_A(x,t)}{dt} = (\frac{\partial f_S(x,y,t)}{\partial x})^T p(t), \qquad \tilde{g}_A(x,t_1) = 0. \tag{$\overline{\text{D2}}$}$$

Totally, $n_S + 1$ differential equations are solved in the forward direction and $2n_S + n$ differential equations are solved in the backward direction.

7.3 Forward computation of the Hessian matrix

Denote $v(x,t) = du(x,t)/dx = d^2y(x,t)/dx^2$, so that $v : R^n \times [t_0, t_1] \to R^{n_S \times n \times n}$. Differentiating ($\overline{\text{O1}}$) and ($\overline{\text{D1}}$) we obtain

$$G(x) = G_A(x,t_1) + u^T(x,t_1) \frac{\partial^2 f_T(y(x,t_1))}{\partial y^2} u(x,t_1) + \frac{\partial f_T(y(x,t_1))}{\partial y} v(x,t_1), \tag{$\overline{\text{O3}}$}$$

where

$$\begin{aligned}\frac{dv(x,t)}{dt} &= \frac{\partial f_S(x,y,t)}{\partial y}v(x,t) \\ &\quad+\left[\frac{\partial^2 f_S(x,y,t)}{\partial y^2}u(x,t)+\frac{\partial^2 f_S(x,y,t)}{\partial y \partial x}\right]u(x,t) \\ &\quad+\frac{\partial^2 f_S(x,y,t)}{\partial x \partial y}u(x,t)+\frac{\partial^2 f_S(x,y,t)}{\partial x^2},\end{aligned}$$

$$v(x,t_0)=\frac{d^2 f_I(x)}{dx^2},$$

$$\frac{dG_A(x,t)}{dt}=u^T(x,t)\frac{\partial^2 f_A(y,t)}{\partial y^2}u(x,t)+\frac{\partial f_A(y,t)}{\partial y}v(x,t), \qquad G_A(x,t_0)=0. \tag{$\overline{\text{D3}}$}$$

Here $G(x)=d^2F(x)/dx^2$ and $G_A(x)=d^2f_A(x,t)/dx^2$. Totally, $(n_S+1)(n^2+n+1)$ differential equations are solved in the forward direction.

7.4 Forward approximation of the Hessian matrix (sum of squares)

$$f_A(y,t)=\frac{1}{2}(y(x,t)-z(t))^T W(t)(y(x,t)-z(t)),$$

$$\frac{\partial f_A(y,t)}{\partial y}=W(t)(y(x,t)-z(t)), \qquad \frac{\partial^2 f_A(y,t)}{\partial y^2}=W(t)$$

and similarly

$$f_T(y(x,t_1))=\frac{1}{2}(y(x,t_1)-z(t_1))^T W_1(y(x,t_1)-z(t_1)),$$

$$\frac{\partial f_T(y(x,t_1))}{\partial y}=W_1(y(x,t_1)-z(t_1)), \qquad \frac{\partial^2 f_T(y(x,t_1))}{\partial y^2}=W_1.$$

Here $z:[t_0,t_1]\to R^{n_S}$, $W:[t_0,t_1]\to R^{n_S\times n_S}$ (SPD) ($W_1\neq W(t_1)$ in general). If $F(x)\to 0$, then necessarily $y(x,t)\to z(t)$, so that $\partial f_A(y(x,t),t)/\partial y\to 0$ and $\partial f_T(y(x,t_1))/\partial y\to 0$. Therefore, these terms can be omitted in ($\overline{\text{O3}}$) and ($\overline{\text{D3}}$). In this way we obtain

$$G(x)\approx B(x)=B_A(x,t_1)+u^T(x,t_1)W_1u(x,t_1), \tag{$\overline{\text{O4}}$}$$

where

$$\frac{dB_A(x,t)}{dt}=u^T(x,t)W(t)u(x,t), \qquad B_A(x,t_0)=0. \tag{$\overline{\text{D4}}$}$$

Totally, $(n_S+1)(n+1)+n^2$ differential equations are solved in the forward direction.

List of the most important notations

Notation	pages
F1-F5	2
S1-S3	8
CG	18, 26, 47
VM1-VM3	27
H	29
HD,HB,HR	29
B	30
BD,BB,BR	31
S	31
SD,SB,SR	33
A	34
AD,AB,AR	34
T1-T3	40
NE,OE,SE	54-55
R1-R2	58
N,ND,NB	58-59
$\overline{\mathrm{H}}$	59
$\overline{\mathrm{HD}}, \overline{\mathrm{HB}}, \overline{\mathrm{HR}}$	60-61
$\overline{\mathrm{BD}}, \overline{\mathrm{BB}}, \overline{\mathrm{BR}}$	62
SL,$\overline{\mathrm{SL}}$	80
BL,$\overline{\mathrm{BL}}$	83
J3-J5	88
A3-A4	89
$\overline{\mathrm{S1}} - \overline{\mathrm{S2}}$	90, 92
$\overline{\mathrm{T1}} - \overline{\mathrm{T4}}$	93-94
D	99
QN1-QN3	100
$\overline{\mathrm{A}}$	100
$\overline{\mathrm{AG}}, \overline{\mathrm{AB}}, \overline{\mathrm{AD}}$	100
$\overline{\mathrm{S}}$	101
$\overline{\mathrm{SG}}, \overline{\mathrm{SB}}, \overline{\mathrm{SI}}$	101
$\overline{\mathrm{AA}}$	105
$\overline{\mathrm{SS}}$	105
$\overline{\mathrm{AS}}$	107
$\overline{\mathrm{AD}}, \overline{\mathrm{AR}}$	110
M	117
$\overline{\mathrm{M}}$	118
O1-O4,D1-D4	124-126

Contents