Inexact Trust Region Methods Based on Preconditioned Iterative Subalgorithms for Large Sparse Systems of Nonlinear Equations

Lukšan, Ladislav
1996

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

Inexact Trust Region Methods Based on Preconditioned Iterative Subalgorithms for Large Sparse Systems of Nonlinear Equations

Ladislav Lukšan Jan Vlček

Technical report No. V-667

March 1996

Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic
phone: (+4202) 66053260 fax: (+4202) 8585789
e-mail: luksan@uivt.cas.cz

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

Inexact Trust Region Methods Based on Preconditioned Iterative Subalgorithms for Large Sparse Systems of Nonlinear Equations

Ladislav Lukšan[1] Jan Vlček

Technical report No. V-667
March 1996

Abstract

This paper is devoted to globally convergent methods for solving large sparse systems of nonlinear equations with an inexact approximation of the Jacobian matrix. These methods include difference versions of the Newton method and various quasi-Newton methods. We propose a class of trust region methods together with a proof of their global convergence and describe an implementable globally convergent algorithm which can be used as a realization of these methods. Considerable attention is concentrated on the application of conjugate gradient-type iterative methods to the solution of linear subproblems. We prove that both the GMRES and the smoothed CGS well-preconditioned methods can be used for the construction of globally convergent trust region methods. The efficiency of our algorithm is demonstrated computationally by using a large collection of sparse test problems.

Keywords

Nonlinear equations,trust region methods, global convergence, inexact Jacobians, nonsymmetric linear systems, conjugate gradient-type methods, residual smoothing

AMS classification. 62J02

[1]This work was supported by the Grant Agency of the Czech Republic under grant 201/96/0918

1 Introduction

Let f be a continuously differentiable mapping from $\mathcal{R}^n$ to $\mathcal{R}^n$ in the form $f(x) = (f_1(x), f_2(x), \ldots, f_n(x))^T$ and consider the system of nonlinear equations

$$f(x) = 0 \tag{1.1}$$

for some unknown point $x \in \mathcal{R}^n$. Let $J(x)$ denote the Jacobian matrix of the mapping f with

$$(J(x))_{ij} = \frac{\partial f_i(x)}{\partial x_j}, 1 \leq i \leq n, 1 \leq j \leq n.$$

Let $x_1 \in \mathcal{R}^n$, $\overline{F} \geq \|f(x_1)\|$ and $\overline{\Delta} > 0$. Denote

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

and

$$\mathcal{D}(\overline{F}, \overline{\Delta}) = \{x \in \mathcal{R}^n : \|x - y\| \leq \overline{\Delta} \text{ for some } y \in \mathcal{L}(\overline{F})\}.$$

Throughout the paper we will use the Euclidean vector norm and the spectral matrix norm and will suppose that the following assumptions hold:

A1: The Jacobian matrix $J(x)$ is defined and bounded on $\mathcal{D}(\overline{F}, \overline{\Delta})$, i.e.

$$\|J(x)\| \leq \overline{J}, \quad \forall x \in \mathcal{D}(\overline{F}, \overline{\Delta}).$$

A2: The Jacobian matrix $J(x)$ is Lipschitz continuous on $\mathcal{D}(\overline{F}, \overline{\Delta})$, i.e.

$$\|J(y) - J(x)\| \leq \overline{L}\|y - x\| \quad \forall x, y \in \mathcal{D}(\overline{F}, \overline{\Delta}).$$

In this paper, we will concentrate on a class of trust region methods for the solution of a system (1.1), which generate a sequence of points $x_i \in \mathcal{R}^n$, $i \in \mathcal{N}$, so that

$$x_{i+1} = x_i + \alpha_i s_i, \quad i \in \mathcal{N}, \tag{1.2}$$

where $s_i \in \mathcal{R}^n$, $\|s_i\| \leq \Delta_i$ is the direction vector determined as to be an inexact minimizer of $\|A_i s + f_i\|$ over the trust region with the radius Δ_i, and where the stepsize α_i is selected so that either $\alpha_i = 1$, if $\|f(x_i + s_i)\| < \|f_i\|$, or $\alpha_i = 0$, otherwise. Here A_i is an approximation of the matrix $J_i = J(x_i)$ and $f_i = f(x_i)$.

For the investigation of trust region methods we also use the objective function

$$F(x) = \frac{1}{2}\|f(x)\|^2, \tag{1.3}$$

which has the same local and global minima as the norm $\|f(x)\|$, and denote $F_i = F(x_i)$, $g_i = g(x_i)$, $i \in \mathcal{N}$, where $g(x) = J^T(x)f(x)$ is the gradient of $F(x)$.

While the influence of inexactness of the solution of the system $A_i s + f_i = 0$ on global convergence was succesfully studied in Refs.1-5, the influence of inexactness of the approximation A_i of the Jacobian matrix J_i has not been considered. Therefore, we consider both of these inexactnesses in this paper.

The paper is organized as follows. In Section 2, we propose a class of truncated trust region methods for nonlinear equations and formulate conditions for their global convergence. These conditions (especially condition (2.9)) cannot be verified in general, but our theory can be useful for particular algorithmic realizations. We introduce an implementable algorithm, based on restarts, which does not use condition (2.9) while it is still globally convergent (if standard assumptions hold). Section 3 is devoted to the investigation of preconditioned iterative methods for the solution of nonsymmetric systems of linear equations. We prove that both the GMRES and the smoothed CGS methods can be used for the construction of globally convergent trust region methods if certain conditions hold. Condition (3.4) is essential, but we propose additional conditions and a rule based on one of them (condition (3.2) cannot be verified in practice, but its significance is in that it is frequently satisfied when a good precondition technique is used). Finally, section 4 contains computational experience with truncated trust region methods that utilize incomplete LU decomposition as a preconditioning technique.

Remark to the notation: Throughout the paper we denote $L_i(s) = \|A_i s + f_i\| - \|f_i\|$ for the predicted decrease of the residual norm and $\rho_i(s) = (\|f(x_i+s)\| - \|f(x_i)\|)/L_i(s)$ for the ratio of both the actual and the predicted decreases of the residual norm.

2 Trust region methods

We begin with the definition of a class of trust region methods for the solution of a system of nonlinear equations. More detailed information can be found in Algorithm 1.

Definition 1 *We say that the basic method $x_{i+1} = x_i + \alpha_i s_i$, $i \in \mathcal{N}$ for the solution of a system of nonlinear equations $f(x) = 0$ is a trust region method (T) if the following conditions hold.*
T1: Direction vectors $s_i \in \mathcal{R}^n$, $i \in \mathcal{N}$, are determined so that

$$\|s_i\| \le \Delta_i, \tag{2.1}$$
$$\|s_i\| < \Delta_i \Rightarrow \|A_i s_i + f_i\| \le \overline{\omega}\|f_i\|, \tag{2.2}$$
$$-L_i(s_i) \ge 2\underline{\sigma}\|A_i s_i\|, \tag{2.3}$$

where $0 \le \overline{\omega} < 1$, $0 < \underline{\sigma} \le 1/2$.
T2: Steplengths $\alpha_i \ge 0$, $i \in \mathcal{N}$, are chosen so that

$$\rho_i(s_i) \le 0 \Rightarrow \alpha_i = 0, \tag{2.4}$$
$$\rho_i(s_i) > 0 \Rightarrow \alpha_i = 1, \tag{2.5}$$

T3: Trust region radii $0 < \Delta_i \le \overline{\Delta}$, $i \in \mathcal{N}$, are chosen so that Δ_1 is given and

$$\rho_i(s_i) < \underline{\rho} \Rightarrow \underline{\beta}\|s_i\| \le \Delta_{i+1} \le \overline{\beta}\|s_i\|, \tag{2.6}$$
$$\rho_i(s_i) \ge \underline{\rho} \Rightarrow \Delta_i \le \Delta_{i+1} \le \overline{\Delta}, \tag{2.7}$$

where $0 < \underline{\beta} \le \overline{\beta} < 1$ *and* $0 < \underline{\rho} < 1/2$.

The constant $\underline{\sigma}$ depends on a particular procedure for direction determination. If the matrix A_i is nonsingular, then a vector s_i satisfying T1 for an arbitrary value $0 < \underline{\sigma} \le 1/2$ exists. Indeed, if we choose $s_i = -\mu_i A_i^{-1} f_i$, where $0 < \mu_i \le 1$ is the maximum number such that $\|s_i\| \le \Delta_i$, then (2.1) and (2.2) hold and, moreover

$$\begin{aligned} -L_i(s_i) &= \|f_i\| - \|A_i s_i + f_i\| = \|f_i\| - \|(1-\mu_i) f_i\| \\ &= \mu_i \|f_i\| = \mu_i \|A_i A_i^{-1} f_i\| = \|A_i s_i\| \ge 2\underline{\sigma} \|A_i s_i\|. \end{aligned}$$

The constants $\overline{\omega}$, $\underline{\beta}$, $\overline{\beta}$ and $\underline{\rho}$ are user supplied and T1 - T3 can always be satisfied for an arbitrary $0 \le \overline{\omega} < 1$, $0 < \underline{\beta} \le \overline{\beta} < 1$ and $0 < \underline{\rho} < 1/2$.

In subsequent considerations, we denote $\mathcal{N}_1$, $\mathcal{N}_2$, $\mathcal{N}_3$ the sets of indices such that $\|s_i\| < \Delta_i$, $\rho_i(s_i) > 0$, $\rho_i(s_i) \ge \underline{\rho}$ holds, respectively. Furthermore, we alternatively use the following assumptions:

A3: Matrices $J_i^{-1} = J^{-1}(x_i)$ *are defined and uniformly bounded on the sequence of points* $x_i \in \mathcal{L}(\overline{F})$, $i \in \mathcal{N}$, *generated by the trust region method (T), i.e.*

$$\|J_i^{-1}\| \le 1/\underline{J}, \quad \forall i \in \mathcal{N}.$$

A4: There exist values $0 < \underline{A} \le \overline{A}$, *such that*

$$\underline{A}\|s_i\| \le \|A_i s_i\| \le \overline{A}\|s_i\| \tag{2.8}$$

for all $i \in N$, *and values* $\underline{J} > 0$ *and* $0 < \overline{\vartheta} < \underline{J}\gamma/(1+\gamma)$, *where* $\gamma = (1-2\underline{\rho})\underline{\sigma} > 0$, *such that* $\|J_i s_i\| \ge \underline{J}\|s_i\|$ *and*

$$\|(A_i - J_i)s_i\| \le \overline{\vartheta}\|s_i\| \tag{2.9}$$

for all $i \in \mathcal{N} \backslash (\mathcal{N}_1 \cup \mathcal{N}_3)$.

Now we will prove that trust region method (T) is globally convergent if A1, A2 and A4 hold. Our proof is motivated by the proof proposed in Ref.6 for unconstrained optimization.

Lemma 1 *Let trust region method (T) be applied to the function* $f : \mathcal{R}^n \to \mathcal{R}^n$ *satisfying assumptions A1 - A2 and let A4 hold. Then a constant* $\underline{c} > 0$ *exists, such that*

$$\|s_i\| \ge \underline{c}\|f_i\|$$

for all $i \in \mathcal{N}$.

Proof (a) Let $i \in \mathcal{N}_1$. Then from (2.2) follows

$$|\|A_i s_i\| - \|f_i\|| \le \|A_i s_i + f_i\| \le \overline{\omega}\|f_i\|,$$

so that $(1-\overline{\omega})\|f_i\| \le \|A_i s_i\|$, which together with (2.8) gives

$$\|s_i\| \ge \frac{1-\overline{\omega}}{\overline{A}}\|f_i\|.$$

(b) Let $i \notin \mathcal{N}_1$ and $i \notin \mathcal{N}_3$. From (2.3) it follows that $L_i(s_i) \leq 0$, so that

$$\begin{aligned} L_i(s_i)\|f_i\| &= (\|A_i s_i + f_i\| - \|f_i\|)\,\|f_i\| \geq \left(\|A_i s_i + f_i\|^2 - \|f_i\|^2\right) \\ &= 2\left(f_i^T A_i s_i + \frac{1}{2} s_i^T A_i^T A_i s_i\right) \triangleq 2Q_i(s_i) \end{aligned} \tag{2.10}$$

If $\|f(x_i + s_i)\| \leq \|f(x_i)\|$, then the inequality $\rho_i(s_i) < \underline{\rho}$ and (2.10) imply

$$\begin{aligned} F(x_i + s_i) - F(x_i) &= \frac{1}{2}\left(\|f(x_i + s_i)\|^2 - \|f(x_i)\|^2\right) \\ &\geq (\|f(x_i + s_i)\| - \|f(x_i)\|)\,\|f(x_i)\| \\ &\geq \underline{\rho} L_i(s_i)\|f_i\| \geq 2\underline{\rho} Q_i(s_i). \end{aligned}$$

If $\|f(x_i + s_i)\| \geq \|f(x_i)\|$, this inequality holds trivially. Therefore we can write

$$F(x_i + s_i) - F(x_i) \geq 2\underline{\rho} Q_i(s_i). \tag{2.11}$$

On the other hand, assumptions A1, A2 and the mean value theorem imply

$$\begin{aligned} \|g(x_i + \mu s_i) - g(x_i)\| &= \|J^T(x_i + \mu s_i) f(x_i + \mu s_i) - J^T(x_i) f(x_i)\| \\ &\leq \|J^T(x_i + \mu s_i)(f(x_i + \mu s_i) - f(x_i))\| \\ &\quad + \|(J^T(x_i + \mu s_i) - J^T(x_i)) f(x_i)\| \\ &\leq \overline{J}\|(f(x_i + \mu s_i) - f(x_i)\| + \overline{L}\mu\|s_i\|\|f_i\| \\ &= \overline{J}\| \int_0^1 J(x_i + \tau\mu s_i)\mu s_i d\tau\| + \overline{L}\mu\|s_i\|\|f_i\| \\ &\leq (\overline{J}^2 + \overline{LF})\|s_i\| \end{aligned}$$

for $0 \leq \mu \leq 1$, so that

$$\begin{aligned} F(x_i + s_i) - F(x_i) &\leq g_i^T s_i + \|g(x_i + \mu s_i) - g(x_i)\|\|s_i\| \\ &\leq g_i^T s_i + (\overline{J}^2 + \overline{LF})\|s_i\|^2 \\ &= f_i^T A_i s_i + f_i^T (J_i - A_i) s_i + (\overline{J}^2 + \overline{LF})\|s_i\|^2 \\ &\leq Q_i(s_i) + \overline{\vartheta}\|s_i\|\|f_i\| + (\overline{J}^2 + \overline{LF})\|s_i\|^2, \end{aligned} \tag{2.12}$$

Coupling (2.11) and (2.12), we can write

$$2\underline{\rho} Q_i(s_i) \leq Q_i(s_i) + \overline{\vartheta}\|s_i\|\|f_i\| + (\overline{J}^2 + \overline{LF})\|s_i\|^2$$

or

$$-(1 - 2\underline{\rho}) Q_i(s_i) \leq \overline{\vartheta}\|s_i\|\|f_i\| + (\overline{J}^2 + \overline{LF})\|s_i\|^2 \tag{2.13}$$

Since $\overline{\vartheta} < \underline{J}\gamma/(1 + \gamma) < \underline{J}$ we can write

$$(\underline{J} - \overline{\vartheta})\|s_i\| \leq \|J_i s_i\| - \|(A_i - J_i)s_i\| \leq \|A_i s_i\|, \tag{2.14}$$

which together with (2.10), and (2.3) imply

$$-Q_i(s_i) \geq -\frac{1}{2} L_i(s_i)\|f_i\| \geq \underline{\sigma}\|A_i s_i\|\|f_i\| \geq \underline{\sigma}(\underline{J} - \overline{\vartheta})\|s_i\|\|f_i\|.$$

If we substitute the last inequality into (2.13), we obtain

$$(1-2\underline{\rho})\underline{\sigma}(\underline{J}-\overline{\vartheta})\|s_i\|\|f_i\| \leq -(1-2\underline{\rho})Q_i(s_i) \leq \overline{\vartheta}\|s_i\|\|f_i\| + (\overline{J}^2+\overline{LF})\|s_i\|^2$$

or

$$\|s_i\| \geq \frac{(1-2\underline{\rho})\underline{\sigma}(\underline{J}-\overline{\vartheta})-\overline{\vartheta}}{\overline{J}^2+\overline{LF}}\|f_i\|$$

since $\|s_i\| \neq 0$ by T1. The numerator is positive since $\overline{\vartheta} < \underline{J}(1-2\underline{\rho})\underline{\sigma}/(1+(1-2\underline{\rho})\underline{\sigma})$. (c) Let $i=1$. If $\|f_1\|=0$, then clearly $\|s_1\| \geq \underline{c}\|f_1\|$ for an arbitrary constant $\underline{c}>0$. If $\|f_1\| \neq 0$, we obtain

$$\|s_1\| \geq \frac{\|s_1\|}{\|f_1\|}\|f_1\|.$$

(d) Let $i \notin \mathcal{N}_1$, $i \in \mathcal{N}_3$ and $i \neq 1$. Let $k<i$ be the maximum index for which $k \notin \mathcal{N}_1$, $k \in \mathcal{N}_3$ and $k \neq 1$ do not hold simulteneously. Then using (2.6), (2.7) and (2.1) 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 we obtain from parts (a)-(c) of this proof (the sequence $\|f_i\|$, $i \in \mathcal{N}$, is nonincreasing by T2)

$$\|s_i\| \geq \underline{\beta}\|s_k\| \geq \underline{c}\|f_k\| \geq \underline{c}\|f_i\|,$$

where

$$\underline{c} = \underline{\beta}\min\left(\frac{1-\overline{\omega}}{\overline{A}}, \frac{(1-2\underline{\rho})\underline{\sigma}(\underline{J}-\overline{\vartheta})-\overline{\vartheta}}{\overline{J}^2+\overline{LF}}, \frac{\|s_1\|}{\|f_1\|}\right). \tag{2.15}$$

□

Theorem 1 *Let $x_i \in \mathcal{R}^n$, $i \in \mathcal{N}$ be a sequence generated by trust region method (T). Let the function $f: \mathcal{R}^n \to \mathcal{R}^n$ satisfy the assumptions A1 - A2 and let A4 hold. Then $x_i \to x^\star$ and $f(x^\star)=0$.*

Proof (a) First we prove that $f_i \to 0$. Suppose that this assertion does not hold. Since the sequence $\|f_i\|$, $i \in \mathcal{N}$, is nonincreasing by T2, a number $\underline{\varepsilon}>0$ exists, such that $\|f_i\| \geq \underline{\varepsilon}$, $\forall i \in \mathcal{N}$ and by Lemma 1 it holds that

$$\|s_i\| \geq \underline{c}\underline{\varepsilon}, \quad \forall i \in \mathcal{N}.$$

Suppose first, that the set $\mathcal{N}_3$ is not finite. Since $\mathcal{N}_3 \subset \mathcal{N}_2$, we can write

$$\begin{aligned} \|f_i\| - \|f_{i+1}\| &= \|f(x_i)\| - \|f(x_i+s_i)\| \geq -\underline{\rho}L_i(s_i) \\ &\geq 2\underline{\rho}\underline{\sigma}\|A_i s_i\| \geq 2\underline{\rho}\underline{\sigma}\underline{A}\underline{c}\underline{\varepsilon}, \ \forall i \in \mathcal{N}_3 \end{aligned} \tag{2.16}$$

by (2.8) and (2.3). Consequently, it follows

$$\begin{aligned}\|f_1\| &\geq \lim_{i\to\infty}(\|f_1\|-\|f_{i+1}\|)=\sum_{i=1}^{\infty}(\|f_i\|-\|f_{i+1}\|)\\ &\geq \sum_{i\in\mathcal{N}_3}(\|f_i\|-\|f_{i+1}\|)\geq\sum_{i\in\mathcal{N}_3}2\underline{\rho}\underline{\sigma}Ac\underline{\varepsilon}=\infty\end{aligned}$$

which is a contradiction. Suppose now, that the set $\mathcal{N}_3$ is finite. Then (2.6) implies $\Delta_i\to 0$, which together with (2.1) gives $\|s_i\|\to 0$. But this is in contradiction to $\|s_i\|\geq\underline{c\varepsilon}\ \forall i\in\mathcal{N}$.
(b) Using (2.3) we obtain $L_i(s_i)=\|A_is_i+f_i\|-\|f_i\|\leq 0$, so that

$$\|f_i\|\geq\|A_is_i+f_i\|\geq\|A_is_i\|-\|f_i\|$$

This inequality implies $\|A_is_i\|\leq 2\|f_i\|$, so that

$$\underline{A}\|s_i\|\leq\|A_is_i\|\leq 2\|f_i\| \tag{2.17}$$

Now we will show that $\sum_{i=1}^{\infty}\|s_i\|<\infty$. If the set $\mathcal{N}_3$ is finite, then an index $l\notin\mathcal{N}_3$ exists, such that $i\notin\mathcal{N}_3\ \forall i\geq l$. Therefore

$$\sum_{i=1}^{\infty}\|s_i\|\leq\sum_{i=1}^{l-1}\|s_i\|+\|s_l\|\sum_{i=l}^{\infty}\overline{\beta}^{i-l}\leq(l-1)\overline{\Delta}+\|s_l\|/(1-\overline{\beta})<\infty$$

by (2.6). If the set $\mathcal{N}_3$ is infinite, then (2.16) implies

$$\begin{aligned}\|f_1\| &\geq \sum_{i=1}^{\infty}(\|f_i\|-\|f_{i+1}\|)\geq\sum_{i\in\mathcal{N}_3}(\|f_i\|-\|f_{i+1}\|)\\ &\geq 2\underline{\rho}\underline{\sigma}\sum_{i\in\mathcal{N}_3}\|A_is_i\|\geq 2\underline{\rho}\underline{\sigma}\underline{A}\sum_{i\in\mathcal{N}_3}\|s_i\|.\end{aligned}$$

Denote $\mathcal{N}_3=\{l_1,l_2,l_3,\ldots\}$. Using Lemma 1 and (2.17), we obtain

$$\|s_{l_j+1}\|\leq\frac{2}{\underline{A}}\|f_{l_j+1}\|\leq\frac{2}{\underline{A}}\|f_{l_j}\|\leq\frac{2}{\underline{cA}}\|s_{l_j}\| \tag{2.18}$$

and (2.6) 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

$$\begin{aligned}\sum_{i=1}^{\infty}\|s_i\| &= \sum_{i=1}^{l_1-1}\|s_i\|+\sum_{j=1}^{\infty}\left[\|s_{l_j}\|+\sum_{k=1}^{l_{j+1}-l_j-1}\|s_{l_j+k}\|\right]\\ &\leq (l_1-1)\overline{\Delta}+\sum_{j=1}^{\infty}\|s_{l_j}\|\left[1+\frac{2}{\underline{cA}}\sum_{k=1}^{l_{j+1}-l_j-1}\overline{\beta}^{k-1}\right]\\ &\leq (l_1-1)\overline{\Delta}+\left[1+\frac{2}{\underline{cA}}\frac{1}{1-\overline{\beta}}\right]\sum_{i\in\mathcal{N}_3}\|s_i\|\\ &\leq (l_1-1)\overline{\Delta}+\left[1+\frac{2}{\underline{cA}}\frac{1}{1-\overline{\beta}}\right]\frac{\|f_1\|}{2\underline{\rho}\underline{\sigma}\underline{A}}<\infty.\end{aligned}$$

From $\sum_{i=1}^{\infty} \|x_{i+1} - x_i\| \leq \sum_{i=1}^{\infty} \|s_i\| < \infty$ it follows that the sequence x_i, $i \in \mathcal{N}$, satisfies the Cauchy condition, so that $x_i \to x^\star$, which together with $f_i \to 0$ gives $f(x^\star) = 0$. □

As the application of the general theory of trust region methods, we will investigate a difference version of the Newton method first.

Lemma 2 *Let assumption A2 be satisfied and let $A(x)$ be a matrix obtained by numerical differentiation, such that*

$$A(x)e_j = \frac{f(x + \delta e_j) - f(x)}{\delta}, \quad 1 \leq j \leq n, \tag{2.19}$$

where e_j, $1 \leq j \leq n$, are columns of the unit matrix of order n and δ is a positive constant. Then

$$\|A(x) - J(x)\| \leq \frac{1}{2}\overline{L}\sqrt{n}\delta.$$

Proof See Ref.3, Lemma 4.2.1. □

Corollary 1 *Let assumptions A1 - A3 be satisfied. Let $A_i \approx J(x_i)$, $i \in \mathcal{N}$, be matrices determined by (2.19) where*

$$\delta < \frac{2\underline{J}\gamma}{\overline{L}\sqrt{n}(1+\gamma)}. \tag{2.20}$$

with $\gamma = (1 - 2\underline{\rho})\underline{\sigma}$. Then A4 holds with $\underline{A} = \underline{J} - \overline{\vartheta}$, $\overline{A} = \overline{J} + \overline{\vartheta}$ and $\overline{\vartheta} < \underline{J}\gamma/(1+\gamma)$.

Proof From Lemma 2 and (2.20) we obtain

$$\|A_i - J_i\| \leq \overline{\vartheta} \triangleq \frac{1}{2}\overline{L}\sqrt{n}\delta < \underline{J}\gamma/(1+\gamma).$$

The rest of assertion follows from the equality $A_i s_i = J_i s_i + (A_i - J_i)s_i$. □

Corollary 1 shows that it is possible to choose number $\delta > 0$, so that the trust region method with difference approximation of the Jacobian matrix, determined by (2.19), is globally convergent. This is true only if all computations are perfect. In the opposite case, round-off errors can destroy this property. We should investigate, when the actual difference, derived from machine precision, satisfies condition (2.20). Nevertheless, we omit this investigation since inequality (2.20) is usually unnecessarily strong and difference versions of the Newton method are very robust in practice, as it is demonstrated in Section 4.

Now we will concentrate our attention to quasi-Newton methods. We will not study these methods in detail (we refer to Refs.3 and 5 for theoretical investigations). Instead, we propose an implementable globally convergent algorithm, which can be used as a realization of an arbitrary quasi-Newton method. Since condition (2.9) cannot be verified when the true Jacobian matrix is not known, we have to use a different approach based on restarts. As we can see, condition (2.9) is used only in step (b) of the proof of Lemma 1 and it can be replaced by the condition $\|s_i\| \geq \underline{\varepsilon}\|f_i\|$, where $\underline{\varepsilon}$ is a suitable constant, in this case. Therefore, if we apply the decision

T4: If $i \in \mathcal{N} \setminus (\mathcal{N}_1 \cup \mathcal{N}_3)$ and $\|s_i\| < \underline{\varepsilon}\|f_i\|$ and $A_i \neq J_i$, then set $A_i = J_i$ and repeat the iteration,

we obtain a trust region method which is globally convergent if A1 - A3 hold. Unfortunately, the above simple rule is not efficient in practice. We have known, from the computational experience, that quasi-Newton methods have to be restarted more frequently (especially in the sparse case when the numerical differentiation described in Ref.7 is inexpensive). Therefore, we recommend the following algorithm, in which A_i is replaced by J_i whenever $\rho_i(s_i) < \underline{\rho}$.

Algorithm 1

Data: $0 < \underline{\beta} < 1$, $0 < \underline{\rho} < \overline{\rho} < 1 < \overline{\gamma}$, $0 < \overline{\omega} < 1$, $\underline{\varepsilon} > 0$, $\overline{\varepsilon} > 0$, $\overline{\Delta} > 0$, $\overline{j} > 0$.

Step 1: Initiation: Choose an initial point $x_1 \in \mathcal{R}^n$ and compute the vector $f_1 := f(x_1)$. Choose a number $0 < \Delta_1 \leq \overline{\Delta}$ and set $i := 1$ and $k := 1$.

Step 2: Test on convergence: If $\|f_i\| \leq \overline{\varepsilon}$, then terminate the computations (the solution is obtained). Otherwise set $j := 0$.

Step 3: Restart: If $k = 1$, then compute the matrix $J_i := J(x_i)$ and set $A_i := J_i$

Step 4: Direction determination: Set $\overline{\omega}_i := \min(\|f_i\|^{1/2}, 1/i, \overline{\omega})$ and compute the vector $s_i \in \mathcal{R}^n$ satisfying the conditions (2.1) - (2.3) (with $\overline{\omega}_i$ instead of $\overline{\omega}$). This vector can be computed by Rule 1 in the basic case or Rule 2 in the preconditioned case (both of these rules are described in Section 3).

Step 5: Stepsize selection: Set $x_{i+1} := x_i + s_i$, compute $f_{i+1} := f(x_{i+1})$ and determine $\rho_i(s_i)$. If $\rho_i(s_i) < \underline{\rho}$, then go to Step 7. If $\underline{\rho} \leq \rho_i(s_i) \leq \overline{\rho}$, then set $\Delta_{i+1} = \Delta_i$. If $\overline{\rho} < \rho_i(s_i)$ and $\|s_i\| < \Delta_i$, then set $\Delta_{i+1} = \Delta_i$. If $\overline{\rho} < \rho_i(s_i)$ and $\|s_i\| = \Delta_i$, then set $\Delta_{i+1} = \min(\overline{\gamma}\Delta_i, \overline{\Delta})$.

Step 6: Update: Compute the matrix A_{i+1} using the quasi-Newton update, set $i := i + 1$, $k := 0$ and go to Step 2.

Step 7: Decision: (a) If $\|s_i\| = \Delta_i < \underline{\varepsilon}\|f_i\|$ and $k = 0$, then set $k := 1$ and go to Step 3. Otherwise set $\Delta_{i+1} = \underline{\beta}\|s_i\|$.

(b) If $\rho_i(s_i) > 0$, then set $i := i + 1$, $k := 1$ and go to Step 2.

(c) If $j \geq \overline{j}$, then terminate the computations (the algorithm fails). Otherwise set $j := j + 1$, $k := k + 1$ and go to Step 3.

We assume that matrices J_i, $i \in \mathcal{N}$, in Step 3, are computed using differences as in (2.19). This computation is very efficient for sparse systems when the technique described in Ref.7 is applied. The value $\overline{\omega}_i := \min(\|f_i\|^{1/2}, 1/i, \overline{\omega})$ is used instead of $\overline{\omega}$

in Step 4, since this choice guarantees the ultimate superlinear rate of convergence of the Newton method (Ref.4). Step 7a realizes decision T4 which guarantees the global convergence of Algorithm 1 in case A1 - A3 hold (A1 - A3 are standard assumptions which are not verified during the computation but their violation can cause a failure of Algorithm 1 in Step 7c).

3 Iterative solution of linear subproblems

The direction vector $s_i \in \mathcal{R}^n$, $i \in \mathcal{N}$, satisfying the inequality $\|A_i s_i + f_i\| \leq \overline{\omega}\|f_i\|$ is most frequently obtained as an approximate solution of the linear subproblem $A_i s + f_i = 0$ using some iterative method. In order to simplify the notation we omit the outer iteration index i in this section, so that we write A, f, x instead of A_i, f_i, x_i. On the other hand, we use the inner iteration index j for the description of iterative methods for linear subproblems. We return to the outer iteration index i only in Theorem 3.

To satisfy conditions (2.1) - (2.3), we need iterative methods which will terminate after a finite number of steps and generate a sequence of iterates s_j, $j \in \mathcal{N}$, and corresponding residual vectors $r_j \stackrel{\Delta}{=} As_j + f$, $j \in \mathcal{N}$, so that the norms $\|r_j\|$, $j \in \mathcal{N}$, do not increase. This requirement can be fulfilled by a choice of some smoothed (residual minimizing) conjugate gradient-type method. Moreover, since the system matrix A is not always explicitly known and is usually given by the difference formula as in Lemma 2, we consider only the iterative methods which do not involve multiplication by the transpose of the matrix A (transpose-free methods). In this section, we always suppose that $s_1 = 0$, so that $r_1 = f$.

One of the best known and widely used schemes of this type is the GMRES method presented by Saad and Schultz in Ref.8 and given by the following algorithm.

Algorithm 2. Preconditioned GMRES method.

$$
\begin{aligned}
& s_1 = 0,\ r_1 = f,\ \beta_1 q_1 = r_1,\ \|q_1\| = 1 \\
& j = 1, 2, \ldots, n \\
& w_1 = AC^{-1} q_j \\
& \left.\begin{array}{l} \alpha_{kj} = q_k^T w_k \\ w_{k+1} = w_k - \alpha_{kj} q_k \end{array}\right\} k = 1, \ldots, j \\
& \beta_{j+1} q_{j+1} = w_{j+1},\ \|q_{j+1}\| = 1 \\
& Q_j = [q_1, q_2, \ldots, q_j] \\
& H_j = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1j} \\ \beta_2 & \alpha_{22} & \ldots & \alpha_{2j} \\ 0 & \beta_3 & \ldots & \alpha_{3j} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \beta_{j+1} \end{bmatrix} \in \mathcal{R}^{(j+1)\times j} \\
& z_j = \arg\min_{z \in \mathcal{R}^j} \|H_j z + \beta_1 e_1\| \\
& s_{j+1} = C^{-1} Q_j z_j
\end{aligned}
$$

Setting $C = I$, we obtain the basic (non-preconditioned) GMRES method. Very often we consider $C = LU$, where the triangular matrices L and U are obtained from the incomplete LU (ILU) decomposition. We use the notation $B = AC^{-1}$ which simplifies the investigation of both the preconditioned and the non-preconditioned versions of the considered methods. We note that the vectors $q_k \in \mathcal{R}^n$, $k = 1, \ldots, j$ constitute the orthonormal basis of the Krylov subspace

$$\mathcal{K}_j = \text{span}\{f, Bf, \ldots, B^{j-1}f\},$$

and we can write $BQ_j = Q_{j+1}H_j$. Moreover, GMRES is a minimimal residual method, the residual norm $\|r_j\|$ is minimized over the Krylov subspace $\mathcal{K}_j$, i.e.

$$s_{j+1} = C^{-1}\tilde{s}_{j+1} = C^{-1} \arg\min_{\tilde{s} \in \mathcal{K}_j} \|B\tilde{s} + f\|.$$

The sequence of residual norms is non-increasing and the solution of a linear system is obtained after at most n iterations (if the rounding errors do not deteriorate the finite termination of the method). Unfortunately, the method uses long recurrences ($O(j^2)$ operations and storage requirement per iteration step). Since $O(n^2)$ can be too large, the GMRES method is often restarted after $m < n$ iterations. We call this modification the GMRES(m) method.

In this paper we will concentrate on another conjugate gradient-type iterative method, smoothed CGS method, presented in Ref.9 and known as one of the most efficient transpose-free methods based on short recurrences.

Algorithm 3. Preconditioned smoothed CGS method.

$$\begin{aligned}
&s_1 = 0,\ \overline{s}_1 = 0,\ r_1 = f,\ \overline{r}_1 = f,\ p_1 = f,\ u_1 = f \\
&j = 1, 2, \ldots, n \\
&v_j = AC^{-1}p_j,\ \alpha_j = f^T\overline{r}_j / f^T v_j \\
&q_j = u_j - \alpha_j v_j \\
&\overline{s}_{j+1} = \overline{s}_j + \alpha_j C^{-1}(u_j + q_j) \\
&\overline{r}_{j+1} = \overline{r}_j + \alpha_j AC^{-1}(u_j + q_j),\ \beta_j = f^T\overline{r}_{j+1} / f^T\overline{r}_j \\
&u_{j+1} = \overline{r}_{j+1} + \beta_j q_j \\
&p_{j+1} = u_{j+1} + \beta_j(q_j + \beta_j p_j) \\
&[\lambda_j, \mu_j]^T = \arg\min_{[\lambda,\mu]^T \in \mathcal{R}^2} \|\overline{r}_{j+1} + \lambda(r_j - \overline{r}_{j+1}) + \mu v_j\| \\
&s_{j+1} = \overline{s}_{j+1} + \lambda_j(s_j - \overline{s}_{j+1}) + \mu_j C^{-1}p_j \\
&r_{j+1} = \overline{r}_{j+1} + \lambda_j(r_j - \overline{r}_{j+1}) + \mu_j v_j
\end{aligned}$$

Since this method is obtained from the two parameter-minimal residual smoothing of the CGS method (Ref.10), the sequence of residual norms is non-increasing. The smoothed CGS method uses short recurrences ($O(n)$ operations and storage requirements per iteration step), but it can break down if either $f^T\overline{r}_j = 0$ or $f^T v_j = 0$. The solution of a linear system is obtained after at most n iterations (if a breakdown

does not occur and if rounding errors do not deteriorate the finite termination of the method).

Both the above algorithms (and also other algorithms having a finite termination property and based on a two parameter-minimal residual smoothing) can be used for a generation of direction vectors satisfying conditions (2.1) - (2.3). We determine the direction vectors by the following rule, which is a generalization of the rule proposed in Ref.11 for unconstrained optimization.

Rule 1 *Let $s_k \in \mathcal{R}^n$, $k \in \mathcal{N}$, be vectors generated using Algorithm 2 or Algorithm 3 Let $j \in \mathcal{N}$ be the maximum index, such that $\|s_k\| < \Delta$ and $\|r_k\| > \overline{\omega}\|f\|$ for all $k = 1, \ldots, j$, where $0 \leq \overline{\omega} < 1$ and $\Delta > 0$. If $\|r_{j+1}\| \leq \overline{\omega}\|f\|$ and $\|s_{j+1}\| < \Delta$, then we set $s = s_{j+1}$. If $\|s_{j+1}\| \geq \Delta$, then we set $s = s_j + \tau_j(s_{j+1} - s_j)$, where τ_j is chosen so that $\|s\| = \Delta$.*

Direction vectors selected by Rule 1 clearly satisfy conditions (2.1) and (2.2) of Definition 1. In the subsequent text, we prove the validity of condition (2.3), i.e.

$$\|f\| - \|As + f\| \geq 2\underline{\sigma}\|As\|, \tag{3.1}$$

where $\underline{\sigma}$ is a constant. We use the following definition.

Definition 2 *We say that the matrix $B = AC^{-1}$ is well-preconditioned if*

$$\|I - B\| \leq \overline{\nu} \tag{3.2}$$

for an arbitrary $0 \leq \overline{\nu} < 1$.

Lemma 3 *Let the matrix B be well-preconditioned and let $s_{j+1} \in \mathcal{R}^n$, $j = 1, \ldots, n$, be the vectors generated by either the GMRES or the smoothed CGS method. Then*

$$\|f\|^2 - \|r_{j+1}\|^2 \geq \underline{\eta}^2\|f\|^2, \tag{3.3}$$

where $\underline{\eta} = (1 - \overline{\nu})/(1 + \overline{\nu})$.

Proof (a) First we prove that

$$|f^T Bf| \geq \frac{1 - \overline{\nu}}{1 + \overline{\nu}}\|f\|\|Bf\| = \underline{\eta}\|f\|\|Bf\|. \tag{3.4}$$

Using Definition 2 we get

$$\begin{aligned} |f^T Bf| &= |f^T f - f^T(I - B)f| \geq |f^T f| - |f^T(I - B)f| \\ &\geq \|f\|^2 - \|I - B\|\|f\|^2 \geq (1 - \overline{\nu})\|f\|^2 \end{aligned}$$

and

$$\|Bf\| \leq \|f\| + \|I - B\|\|f\| \leq (1 + \overline{\nu})\|f\|.$$

Together these inequalities give (3.4).

(b) Since the residual norms of both the GMRES and the smoothed CGS method does not increase, it suffices to prove, that

$$\|f\|^2 - \|r_2\|^2 \geq \underline{\eta}^2 \|f\|^2.$$

Consider first the GMRES method. Since $s_1 = 0$ and $\mathcal{K}_1 = \mathrm{span}\{f\}$, we obtain

$$\|r_2\| = \min_{\mu \in R} \|B(\mu f) + f\|.$$

From the optimality condition

$$\mu_1 \stackrel{\Delta}{=} \arg\min_{\mu \in R} \|B(\mu f) + f\|^2 = \arg\min_{\mu \in R} (\mu^2 \|Bf\|^2 + 2\mu f^T Bf + \|f\|^2)$$

it follows $\mu_1 = -f^T Bf / \|Bf\|^2$ and for the norm of the residual r_2 we have

$$\|r_2\|^2 = \frac{(f^T Bf)^2}{\|Bf\|^4} \|Bf\|^2 - 2\frac{(f^T Bf)^2}{\|Bf\|^2} + \|f\|^2 = \|f\|^2 - \frac{(f^T Bf)^2}{\|Bf\|^2 \|f\|^2} \|f\|^2.$$

This equality together with (3.4) implies assertion of the lemma for the GMRES method. Consider now the smoothed CGS method. Then it holds

$$\|r_2\| = \min_{[\lambda,\mu]^T \in R^2} \|\overline{r}_2 + \lambda(f - \overline{r}_2) + \mu v_1\| \leq \min_{\mu \in R} \|f + \mu v_1\| = \min_{\mu \in R} \|f + \mu Bf\|$$

(after substituting $\lambda = 1$) and we obtain the same result as for the GMRES method. □

Lemma 4 *Let the assumption of Lemma 3 be satisfied and let $s \in \mathcal{R}^n$ be a vector determined by Rule 1. Then*

$$\|f\| - \|As + f\| \geq 2\underline{\sigma} \|As\|, \tag{3.5}$$

where $2\underline{\sigma} = \underline{\eta}^2/8$.

Proof (a) Let $\|s_{j+1}\| < \Delta$ and $\|r_{j+1}\| \leq \overline{\omega}\|f\|$. Then by Lemma 3

$$2\|f\| (\|f\| - \|r_{j+1}\|) \geq \|f\|^2 - \|r_{j+1}\|^2 \geq \underline{\eta}^2 \|f\|^2$$

holds, which together with (2.17) implies

$$\|f\| - \|r_{j+1}\| \geq \frac{1}{2}\underline{\eta}^2 \|f\| \geq \frac{1}{4}\underline{\eta}^2 \|As\|,$$

and we obtain an assertion of the lemma.
(b) Let $\|s_{j+1}\| \geq \Delta$ and $j > 1$. Then $s = \tau_j s_{j+1} + (1 - \tau_j) s_j$ with $0 < \tau_j \leq 1$ holds, so that

$$\|As + f\| = \|\tau_j (As_{j+1} + f) + (1 - \tau_j)(As_j + f)\| \leq \tau_j \|r_{j+1}\| + (1 - \tau_j)\|r_j\|,$$

so that Lemma 3, together with (2.17), gives

$$\|f\| - \|As + f\| \geq \tau_j \left(\|f\| - \|r_{j+1}\|\right) + (1 - \tau_j)\left(\|f\| - \|r_j\|\right) \geq \frac{1}{2}\underline{\eta}^2\|f\| \geq \frac{1}{4}\underline{\eta}^2\|As\|$$

and we obtain an assertion of the lemma.
(c) Let $\|s_{j+1}\| \geq \Delta$ and $j = 1$. Then $s = \tau_1 s_2$ holds, where $0 < \tau_1 \leq 1$. Therefore, we can write

$$\begin{aligned}
\|f\|^2 - \|As + f\|^2 &= \|f\|^2 - \tau_1^2\|As_2\|^2 - 2\tau_1 f^T As_2 - \|f\|^2 \\
&= -\tau_1^2\|As_2\|^2 - 2\tau_1 f^T As_2 \geq \tau_1\left(-\|As_2\|^2 - 2f^T As_2\right) \\
&= \tau_1\left(\|f\|^2 - \|As_2 + f\|^2\right)
\end{aligned}$$

(since $\tau_1^2 \leq \tau_1$ for $0 < \tau_1 \leq 1$), or

$$\begin{aligned}
2\|f\|\left(\|f\| - \|As + f\|\right) &\geq \|f\|^2 - \|As + f\|^2 \geq \tau_1\left(\|f\|^2 - \|r_2\|^2\right) \\
&\geq \tau_1\|f\|\left(\|f\| - \|r_2\|\right),
\end{aligned}$$

which gives

$$\|f\| - \|As + f\| \geq \frac{1}{2}\tau_1\left(\|f\| - \|r_2\|\right) \geq \frac{1}{4}\tau_1\underline{\eta}^2\|f\|.$$

as in (a). Therefore,

$$2\|f\| \geq \|r_2 - f\| = \|As_2\|$$

which after a substitution into the previous inequality gives

$$\|f\| - \|As + f\| \geq \frac{1}{8}\tau_1\underline{\eta}^2\|As_2\| = \frac{1}{8}\underline{\eta}^2\|As\|$$

and we obtain an assertion of the lemma. □

Theorem 3 *Let matrices $B_i = A_i C_i^{-1}$, $i \in \mathcal{N}$, be well-preconditioned and direction vectors $s_i \in \mathcal{R}^n$, $i \in \mathcal{N}$ be determined by Rule 1. Then conditions (2.1) - (2.3) are satisfied and we can use direction vectors $s_i \in \mathcal{R}^n$, $i \in \mathcal{N}$ for the construction of trust region method (T). If this trust region method is applied to the function $f : \mathcal{R}^n \to \mathcal{R}^n$ satisfying assumptions A1 - A2 and if A3 - A4 hold, then $x_i \to x^\star$ and $f(x^\star) = 0$.*

Proof Assertion of the theorem is an immediate consequence of Lemma 4 and Theorem 1.

□

Condition (3.2) is advantageous in the sense that it depends only on the preconditioning technique. On the other hand, it is rather strong and is verified with difficulty. We can only suppose that (3.2) is satisfied if a good preconditioning technique is applied. This is often the case (at least if problems have band structure) when we utilize the incomplete LU decomposition as a preconditioning technique. Fortunately, we can use the weaker condition

$$\|f - Bf\| \leq \overline{\omega}\|f\|, \tag{3.6}$$

which implies (3.4) and forms a basis for the following rule.

Rule 2 *Let* $\tilde{s} = -C^{-1}f$ *and* $\tilde{r} = A\tilde{s} + f$. *If* $\|\tilde{r}\| \leq \overline{\omega}\|f\|$ *and* $\|\tilde{s}\| < \Delta$, *then we set* $s = \tilde{s}$. *If* $\|\tilde{r}\| \leq \overline{\omega}\|f\|$ *and* $\|\tilde{s}\| \geq \Delta$, *then we set* $s = \tilde{\tau}\tilde{s}$, *where* $\tilde{\tau}$ *is chosen so that* $\|s\| = \Delta$. *If* $\|\tilde{r}\| > \overline{\omega}\|f\|$, *then we use Rule 1.*

It is clear, that (3.5) holds, with $16\sigma = (1-\overline{\omega})^2/(1+\overline{\omega})^2$, if $s = \tilde{s}$ (suffice to follow part (c) of proof of Lemma 4). If $s \neq \tilde{s}$, then we have to suppose the validity of (3.4). We investigated, in our computational experiments, how frequently (3.4) was violated when Rule 2 was applied in connection with the incomplete LU decomposition. We found that the direction vector $s = \tilde{s}$ was used in 90 % of iterations. In the other iterations, condition (3.4) was always satisfied with $\eta = 0.1$.

4 Computational experiments

In this section, we present the results of a comparative study of both the Newton and the Schubert (Ref.12) method, realized as trust region (TR) methods with inexact iterative solution of linear subproblems by either the smoothed CGS (SCGS) or the restarted GMRES (GMRES(m)) methods (m=30 for the basic case and m=10 for the preconditioned case). These methods were implemented using the modular interactive system for universal functional optimization UFO (Ref.13). We used the values $\underline{\beta} = 0.5$, $\underline{\rho} = 0.1$, $\overline{\rho} = 0.9$, $\overline{\gamma} = 2.0$, $\overline{\omega} = 0.4$, $\underline{\varepsilon} = 10^{-8}$, $\overline{\varepsilon} = 10^{-16}$, $\overline{j} = 5$ and $\Delta_1 = 1.0$. These values were obtained experimentally. All test results were obtained using the 17 sparse problems from Ref.14, having 100 equations and 100 unknowns. A summary of the results for all these problems is given in tables 1-2. These tables contain the total number of iterations NIT, the total number of function evaluations NFV, the number of fails, the total computational time and the storage requirement in k-Bytes, for both the basic and the ILU-preconditioned implementations. For comparison we give results obtained by methods with an exact direct solution of linear subproblems by the unsymmetric-pattern multifrontal scheme realized in the UMFPACK package (Ref.15).

Table 1: *Computational variants of the Newton method.*

	Basic (m=30)					Preconditioned (m=10)				
realization	NIT	NFV	Fail	Time	kB	NIT	NFV	Fail	Time	kB
TR-SCGS	382	1641	-	7.90	22	212	968	-	3.52	29
TR-GMRES(m)	285	1349	1	11.97	76	214	980	-	3.72	44
TR-UMFPACK						203	906	-	4.72	71

Table 2: *Computational variants of the Schubert method.*

	Basic (m=30)					Preconditioned (m=10)				
realization	NIT	NFV	Fail	Time	kB	NIT	NFV	Fail	Time	kB
TR-SCGS	635	1123	-	10.27	22	550	804	-	5.44	29
TR-GMRES(m)	504	937	1	17.41	76	547	801	-	5.55	44
TR-UMFPACK						579	927	-	8.51	71

According to the results presented in tables 1-2, we give several conclusions (which are, of course, influenced by our collection of test problems). First, the smoothed CGS method is more efficient than the restarted GMRES method, especially in the basic (unpreconditioned) case. Even if we have not treated possible breakdowns in our implementation of the smoothed CGS method, these breakdowns never occured. Second, the ILU preconditioning substantially improves the efficiency of both the smoothed CGS and the restarted GMRES methods. The improvement obtained by the ILU preconditioning of nonsymmetric iterative methods seems to be more significant then that obtained by the incomplete Choleski preconditioning of the symmetric conjugate gradient method. Third, the preconditioned iterative methods are more efficient than the direct ones measured by both the storage requirements and the total computational time. Finally, we have to note, that we also tested non-preconditioned GMRES method with $m = 10$, but this choice was not efficient (more inner iterations and larger time). Our choices $m = 30$ for the non-preconditioned case and $m = 10$ for the preconditioned case were obtained experimentally.

References

1. BROWN, P.N., SAAD, Y., *Hybrid Krylov Methods for Nonlinear Systems of Equations,* SIAM Journal on Scientific and Statistical Computations, Vol. 11, pp. 450-481, 1990.

2. BROWN, P.N., SAAD, Y., *Convergence Theory of Nonlinear Newton-Krylov Algorithms,* SIAM Journal on Optimization, Vol. 4, pp. 297-330, 1994.

3. DENNIS, J.E., SCHNABEL, R.B., *Numerical Methods for Unconstrained Optimization and Nonlinear Equations,* Prentice-Hall, Englewood Cliffs, New Jersey, 1983.

4. EISENSTAT, S.C., WALKER, H.F., *Globally convergent Inexact Newton Methods,* SIAM Journal on Optimization, Vol. 4, pp. 393-422, 1994.

5. KELLEY, C.T., *Iterative Methods for Linear and Nonlinear Equations,* SIAM, Philadelphia, Pennsylvania, 1995.

6. POWELL, M.J.D., *On the Global Convergence of Trust Region Algorithms for Unconstrained Minimization,* Mathematical Programming, Vol. 29, pp. 297-303, 1984.

7. COLEMAN, T.F., GARBOW, B.S., MORÉ, J.S., *Software for estimating sparse Jacobian matrices,* ACM Transactions of Mathematical Software, Vol. 10, pp. 329-345, 1984.

8. SAAD, Y., SCHULTZ, M., *GMRES a Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems,* SIAM Journal on Scientific and Statistical Computations, Vol. 7, pp. 856-869, 1986.

9. TONG, C.H., *A Comparative Study of Preconditioned Lanczos Methods for Nonsymmetric Linear Systems,* Sandia National Laboratories, Sandia Report No. SAND91-8240B, Livermore, 1992.

10. SONNEVELD, P., *CGS, a Fast Lanczos-Type Solver for Nonsymmetric Linear Systems,* SIAM Journal on Scientific and Statistical Computations, Vol. 10, pp. 36-52, 1989.

11. STEIHAUG, T., *The Conjugate Gradient Method and Trust Regions in Large-Scale Optimization,* SIAM Journal on Numerical Analysis, Vol. 20, pp. 626-637, 1983.

12. SCHUBERT, L.K., *Modification of a Quasi-Newton Method for Nonlinear Equations with Sparse Jacobian,* Mathematics of Computation, Vol. 24, pp. 27-30, 1970.

13. LUKŠAN, L., ŠIŠKA, M., TŮMA, M., VLČEK, J., and RAMEŠOVÁ, N., *Interactive System for Universal Functional Optimization (UFO), Version 1995,* Research Report No. V-662, Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, Czech Republic, 1995.

14. LUKŠAN, L., *Inexact Trust Region Method for Large Sparse Systems of Nonlinear Equations,* Journal of Optimization Theory and Applications, Vol. 81, pp. 569-590, 1994.

15. DAVIS, T.A., *User's Guide for the Unsymmetric-Pattern Multifrontal Package (UMFPACK),* Research Report No. TR-93-020, Computer and Information Sciences Department, University of Florida, Gainesville, Florida, 1993.