Institute of Computer Science
Academy of Sciences of the Czech Republic

A shifted Steihaug-Toint method for computing a trust-region step

Ladislav Lukšan, Ctirad Matonoha, Jan Vlček

Technical report No. 914

16. září 2004

Pod Vodárenskou věží 2, 182 07 Prague 8 phone: +420 2 688 42 44, fax: +420 2 858 57 89,
e-mail:e-mail:ics@cs.cas.cz

Institute of Computer Science
Academy of Sciences of the Czech Republic

A shifted Steihaug-Toint method for computing a trust-region step

Ladislav Lukšan, Ctirad Matonoha, Jan Vlček [1]

Technical report No. 914

16. září 2004

Abstrakt:

Trust-region methods are very convenient in connection with the Newton method for unconstrained optimization. The Moré-Sorensen direct method and the Steihaug-Toint iterative method are most commonly used for solving trust-region subproblems. We propose a method which combines both of these approaches. Using the small-size Lanczos matrix, we apply the Moré-Sorensen method to a small-size trust-region subproblem to compute an approximation of the Lagrange multiplier. Then we solve the shifted system by the Steihaug-Toint method. This paper contains a complete theory concerning properties of the Lagrange multipliers and proves that the new method is globally convergent in the preconditioned case. Finally, results of extensive computational experiments are presented, which demonstrate an efficiency of the new method.

Keywords:
Unconstrained optimization, large-scale optimization, trust-region methods, trust-region subproblems, conjugate gradients, Krylov subspaces, computational experiments.

[1]This work was supported by the Grant Agency of the Czech Academy of Sciences, project code IAA1030405, and by the Ministry of Education of the Czech Republic, project code MSM 242200002.
Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07 Praha 8 and Technical University of Liberec, Hálkova 6, 461 17 Liberec

1 Introduction

Basic optimization methods can be realized in various ways which differ in direction determination and step-size selection. Line-search and trust-region realizations are most popular. Trust-region methods can be advantageously used when the Hessian matrix of the objective function (or its approximation) is indefinite, ill-conditioned or singular. This situation often arises in connection with the Newton method for general objective function (indefiniteness) or with the Gauss-Newton method for nonlinear least-squares (near-singularity). Denote

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

the quadratic function which locally approximates difference $F(x_i + d) - F(x_i)$,

$$\omega_i(d) = (B_i d + g_i)/||g_i||$$

the accuracy of direction determination and

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

the ratio of actual and predicted decrease of the objective function. Trust-region methods generate points $x_i \in \mathcal{R}^n$, $i \in \mathcal{N}$, in such a way that x_1 is arbitrary and

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

where $d_i \in \mathcal{R}^n$ are direction vectors and $\alpha_i > 0$ are step-sizes. Direction vectors $d_i \in \mathcal{R}^n$ are chosen to satisfy conditions

$$\begin{aligned} ||d_i|| &\leq \Delta_i, && (2)\\ ||d_i|| &< \Delta_i \quad \Rightarrow \quad ||\omega_i(d_i)|| \leq \overline{\omega}_i, && (3)\\ -Q_i(d_i) &\geq \underline{\sigma}||g_i|| \min(||d_i||, ||g_i||/||B_i||), && (4) \end{aligned}$$

where $0 \leq \overline{\omega}_i \leq \overline{\omega} < 1$ and $0 < \underline{\sigma} < 1$. Step-sizes $\alpha_i \geq 0$ are selected so that

$$\begin{aligned} \rho_i(d_i) \leq 0 &\Rightarrow \alpha_i = 0, && (5)\\ \rho_i(d_i) > 0 &\Rightarrow \alpha_i = 1. && (6) \end{aligned}$$

Trust-region radii $0 < \Delta_i \leq \overline{\Delta}$ are chosen in such a way that $0 < \Delta_1 \leq \overline{\Delta}$ is arbitrary and

$$\begin{aligned} \rho_i(d_i) < \underline{\rho} &\Rightarrow \underline{\beta}||d_i|| \leq \Delta_{i+1} \leq \overline{\beta}||d_i||, && (7)\\ \rho_i(d_i) \geq \underline{\rho} &\Rightarrow \Delta_i \leq \Delta_{i+1} \leq \overline{\Delta}, && (8) \end{aligned}$$

where $0 < \underline{\beta} \leq \overline{\beta} < 1$ and $0 < \underline{\rho} < 1$. The following theorem, see [7], characterizes the global convergence of trust-region methods.

Theorem 1. *Let objective function $F : \mathcal{R}^N \to \mathcal{R}$ be bounded from below and have bounded second-order derivatives. Consider trust-region method (2)-(8) and denote $M_i = \max(||B_1||, \ldots, ||B_i||)$, $i \in \mathcal{N}$. If*

$$\sum_{i \in \mathcal{N}} \frac{1}{M_i} = \infty, \tag{9}$$

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

Note that (9) is satisfied if there exist constant $\overline{B}$ and infinite set $\mathcal{M} \subset \mathcal{N}$ such that $||B_i|| \leq \overline{B}\ \forall i \in \mathcal{M}$.

A crucial part of each trust region method is a direction determination. We restrict our attention to problems with large dimensions. There are various commonly known methods for computing direction vectors satisfying conditions (2)-(4) which we now mention briefly. To simplify the notation, we omit index i and use symbols $\succeq$ and $\succ$ for ordering by positive semidefiniteness and positive definiteness, respectively.

The most sophisticated method is based on a computation of the optimal locally constrained step. In this case, vector $d \in R^n$ is obtained by solving subproblem

$$\text{minimize} \quad Q(d) = \frac{1}{2} d^T B d + g^T d \quad \text{subject to} \quad ||d|| \leq \Delta. \tag{10}$$

Necessary and sufficient conditions for this solution are

$$||d|| \leq \Delta, \quad (B + \lambda I)d + g = 0, \quad B + \lambda I \succeq 0, \quad \lambda \geq 0, \quad \lambda(\Delta - ||d||) = 0. \tag{11}$$

The Moré-Sorensen method [5] is based on solving nonlinear equation $1/||d(\lambda)|| = 1/\Delta$ with $(B+\lambda I)d(\lambda)+g = 0$ by the Newton method using the sparse Choleski decomposition of $B + \lambda I$. This method is very robust but requires 2-3 Choleski decompositions per iteration.

Simpler methods are based on minimization of $Q(d)$ on the two-dimensional subspace containing Cauchy step $d_C = -(g^T g / g^T B g) g$ and Newton step $d_N = -B^{-1} g$. The most popular is the dog-leg method [6], [2], where $d = d_N$ if $d_N \leq \Delta$ and $d = (\Delta / ||d_C||) d_C$ if $||d_C|| \geq \Delta$. In the remaining case, d is a convex combination of d_C and d_N such that $||d|| = \Delta$. This method requires only one Choleski decomposition per iteration.

If B is not sufficiently sparse, then the sparse Choleski decomposition of B is expensive. In this case, iterative methods based on conjugate gradients are more suitable. Steihaug [8] and Toint [9] proposed a method based on the fact that $Q(d_{k+1}) < Q(d_k)$ and $||d_{k+1}|| > ||d_k||$ hold in the subsequent CG iterations if CG coefficients are positive. We either obtain an unconstrained solution with a sufficient precision or stop on the trust-region boundary if a negative curvature is indicated or the trust-region is left. This method is very efficient in practice especially when suitable preconditioning is used. Note that $||d_{k+1}||_C > ||d_k||_C$ (where $||d_k||_C^2 = d_k^T C d_k$) holds instead of $||d_{k+1}|| > ||d_k||$ if preconditioner C (symmetric and positive definite) is used. Thus the solution on the trust-region boundary obtained by the preconditioned CG method can be farther from the optimal locally constrained step than the solution obtained without preconditioning. This insufficiency is usually compensated by the rapid convergence of the preconditioned CG method.

The CG steps can be combined with Newton step d_N in the multiple dog-leg method [8]. Let $k \ll n$ (usually $k = 5$) and d_k be a vector obtained after k CG steps of the Steihaug-Toint method. If $||d_k|| < \Delta$, we use d_k instead of $d_C = d_1$ in the dog-leg method.

The solution on the trust-region boundary obtained by the Steihaug-Toint method can be rather far from the optimal solution. This insufficiency can be overcame by using the Lanczos process [3]. Initially, the conjugate gradient algorithm is used as in the Steihaug-Toint method. At the same time, the Lanczos tridiagonal matrix is constructed from the CG coefficients. If a negative curvature is indicated or the trust-region is left, we turn to the Lanczos process. In this case, $d = Z\tilde{d}$, where $\tilde{d}$ is obtained by minimizing quadratic function

$$\frac{1}{2} \tilde{d}^T T \tilde{d} + ||g|| e_1^T \tilde{d}$$

subject to $\|\tilde{d}\| \leq \Delta$. Here $T = Z^T B Z$ (with $Z^T Z = I$) is the Lanczos tridiagonal matrix and e_1 is the first column of the unit matrix. This method cannot be successfully preconditioned, since preconditioning changes trust-region constraint $\|d\| \leq \Delta$ to $\|d\|_C \leq \Delta$, where C changes in each major iteration and can be ill-conditioned.

In this paper, we apply the Steihaug-Toint method to subproblem

$$\text{minimize} \quad \tilde{Q}(d) = Q_{\tilde{\lambda}}(d) = \frac{1}{2} d^T (B + \tilde{\lambda} I) d + g^T d \quad \text{subject to} \quad \|d\| \leq \Delta. \tag{12}$$

Number $\tilde{\lambda} \geq 0$, which approximates λ in (11), is found by solving a small-size subproblem with tridiagonal matrix T obtained by using a small number of the Lanczos steps. This method, like method [3], combines good properties of the Moré-Sorensen and the Steihaug-Toint methods. Moreover, it can be successfully preconditioned. The point on the trust-region boundary obtained by this method is usually closer to the optimal solution in comparison with the point obtained by the original Steihaug-Toint method. Section 2 contains theoretical background concerning this method. Computational results are reported in Section 3.

2 A shifted Steihaug-Toint method

A shifted Steihaug-Toint method differs from the standard one by using shifted subproblem (12), where number $\tilde{\lambda}$ approximates λ in (11). Number $\tilde{\lambda}$ has to be chosen in such a way that $\tilde{\lambda} = 0$ if $\|d\| < \Delta$, where d is a solution of (10), which is true if $0 \leq \tilde{\lambda} \leq \lambda$. Thus Theorem 2 allows us to use an arbitrary Krylov subspace for determination of $\tilde{\lambda}$. In this section, we denote by $\mathcal{K}_k = \text{span}\{g, Bg, \ldots, B^{k-1}g\}$ the Krylov subspace of dimension k defined by matrix B and vector g, and by $Z_k \in \mathcal{R}^{n \times k}$ the matrix whose columns form an orthonormal basis in $\mathcal{K}_k$.

Lemma 1. *Let B be symmetric and positive definite matrix, let*

$$\mathcal{K}_j = \text{span}\{g, Bg, \ldots, B^{j-1}g\}, \quad j \in \{1, \ldots, n\},$$

be the j-th Krylov subspace given by matrix B and vector g and let

$$d_j = \arg\min_{d \in \mathcal{K}_j} Q(d), \quad \textit{where} \quad Q(d) = \frac{1}{2} d^T B d + g^T d.$$

If $1 \leq k \leq l \leq n$, then

$$\|d_k\| \leq \|d_l\|.$$

Especially

$$\|d_k\| \leq \|d_n\|, \quad \textit{where} \quad d_n = \arg\min_{d \in \mathcal{R}^n} Q(d).$$

Proof. The assertion of the lemma holds for vectors d_j, $j \geq 1$, generated by the conjugate gradient method starting from $d_0 = 0$ (see [8]). These vectors are minimizers of $Q(d)$ on Krylov subspaces $\mathcal{K}_j$, $j \geq 1$. □

Corollary 1. *Let B be symmetric and positive definite and let $Z_k \in \mathcal{R}^{n \times k}$ be the matrix whose columns form an orthonormal basis in $\mathcal{K}_k$. Then*

$$g^T Z_k (Z_k^T B Z_k)^{-2} Z_k^T g \leq g^T B^{-2} g.$$

Proof. Vector $d_n = -B^{-1}g$ minimizes $Q(d)$ on $\mathcal{R}^n$. Furthermore, if $d = Z_k\tilde{d}$, then

$$Q(d) = Q(Z_k\tilde{d}) = \frac{1}{2}\,\tilde{d}^T Z_k^T B Z_k \tilde{d} + g^T Z_k \tilde{d}.$$

Thus a minimizer of $Q(d)$ on $\mathcal{K}_k$ has form

$$d_k = Z_k\tilde{d}_k = -Z_k(Z_k^T B Z_k)^{-1} Z_k^T g \tag{13}$$

and since $Z_k^T Z_k = I$, Lemma 1 implies that

$$\|d_k\|^2 \leq \|d_n\|^2 \quad \Rightarrow \quad g^T Z_k (Z_k^T B Z_k)^{-2} Z_k^T g \leq g^T B^{-2} g.$$

□

Lemma 2. *Let* $\lambda \in \mathcal{R}$ *and*

$$\mathcal{K}_k(\lambda) = \mathrm{span}\{g, (B+\lambda I)g, \ldots, (B+\lambda I)^{k-1}g\}, \quad k \in \{1, \ldots, n\},$$

be the k*-th Krylov subspace given by matrix* $B + \lambda I$ *and vector* g*. Then*

$$\mathcal{K}_k(\lambda) = \mathcal{K}_k(0). \tag{14}$$

Proof. Equality (14) immediately follows for $k = 1$ because $\mathcal{K}_1(\lambda) = \mathrm{span}\{g\} = \mathcal{K}_1(0)$. Suppose now that (14) holds for some k. Then

$$(B+\lambda I)^k g = (B+\lambda I)(B+\lambda I)^{k-1} g = (B+\lambda I)v = Bv + \lambda v,$$

where $v \in \mathcal{K}_k(\lambda) = \mathcal{K}_k(0)$. As $\lambda v \in \mathcal{K}_k(0)$ and $Bv \in \mathcal{K}_{k+1}(0)$, we have $(B+\lambda I)^k g \in \mathcal{K}_{k+1}(0)$. Thus $\mathcal{K}_{k+1}(\lambda) \subset \mathcal{K}_{k+1}(0)$. Applying the same procedure to matrices $(B+\lambda I)$ and $B = (B+\lambda I) - \lambda I$ we obtain the opposite inclusion. □

Lemma 3. *Let* B_1 *and* B_2 *be symmetric and positive definite matrices. Then*

$$\begin{aligned} B_1 - B_2 \succeq 0 \quad &\textit{if and only if} \quad B_2^{-1} - B_1^{-1} \succeq 0, \\ B_1 - B_2 \succ 0 \quad &\textit{if and only if} \quad B_2^{-1} - B_1^{-1} \succ 0. \end{aligned}$$

Proof. It follows from relations

$$B_1 - B_2 = B_2^{\frac{1}{2}}(B_2^{-\frac{1}{2}} B_1 B_2^{-\frac{1}{2}} - I) B_2^{\frac{1}{2}}, \quad B_2^{-1} - B_1^{-1} = B_1^{-\frac{1}{2}}(B_1^{\frac{1}{2}} B_2^{-1} B_1^{\frac{1}{2}} - I) B_1^{-\frac{1}{2}}$$

and from the fact that matrices $B_2^{-\frac{1}{2}} B_1 B_2^{-\frac{1}{2}}$ and $B_1^{\frac{1}{2}} B_2^{-1} B_1^{\frac{1}{2}}$ have the same eigenvalues. □

Lemma 4. *Let* $Z_k^T B Z_k + \lambda_i I$, $\lambda_i \in \mathcal{R}$, $i \in \{1, 2\}$, *be symmetric and positive definite. Let*

$$d_k(\lambda_i) = \arg\min_{d \in \mathcal{K}_k} Q_{\lambda_i}(d), \quad \textit{where} \quad Q_\lambda(d) = \frac{1}{2}\, d^T(B+\lambda I)d + g^T d.$$

Then

$$\lambda_2 \leq \lambda_1 \quad \Leftrightarrow \quad \|d_k(\lambda_2)\| \geq \|d_k(\lambda_1)\|.$$

Proof. It follows from (13) that

$$\|d_k(\lambda_i)\|^2 = g^T Z_k (Z_k^T (B + \lambda_i I) Z_k)^{-2} Z_k^T g = g^T Z_k (Z_k^T B Z_k + \lambda_i I)^{-2} Z_k^T g$$

with $Z_k^T B Z_k + \lambda_i I$ positive definite. Thus

$$\|d_k(\lambda_2)\|^2 - \|d_k(\lambda_1)\|^2 = g^T Z_k \left[(Z_k^T B Z_k + \lambda_2 I)^{-2} - (Z_k^T B Z_k + \lambda_1 I)^{-2} \right] Z_k^T g.$$

Letting $\tilde{B}_2 = Z_k^T B Z_k + \lambda_2 I$ and assuming that $\lambda_2 \leq \lambda_1$ we can write

$$\begin{aligned} (Z_k^T B Z_k + \lambda_1 I)^2 - (Z_k^T B Z_k + \lambda_2 I)^2 &= (\tilde{B}_2 + (\lambda_1 - \lambda_2) I)^2 - \tilde{B}_2^2 \\ &= 2(\lambda_1 - \lambda_2)\tilde{B}_2 + (\lambda_1 - \lambda_2)^2 I \succeq 0. \end{aligned}$$

Therefore

$$(Z_k^T B Z_k + \lambda_2 I)^{-2} - (Z_k^T B Z_k + \lambda_1 I)^{-2} \succeq 0$$

by Lemma 3, which implies $\|d_k(\lambda_2)\|^2 - \|d_k(\lambda_1)\|^2 \geq 0$. Using the same procedure and the second assertion of Lemma 3 it can be proved that $\lambda_2 < \lambda_1 \Rightarrow \|d_k(\lambda_2)\|^2 > \|d_k(\lambda_1)\|^2$. Since λ_1 and λ_2 can be changed, we obtain $\lambda_1 < \lambda_2 \Rightarrow \|d_k(\lambda_1)\|^2 > \|d_k(\lambda_2)\|^2$ which gives $\|d_k(\lambda_2)\| \geq \|d_k(\lambda_1)\| \Rightarrow \lambda_2 \leq \lambda_1$. □

Theorem 2. *Let d_j, $j \in \{1, \ldots, n\}$, be solutions of minimization problems*

$$d_j = \arg\min_{d \in \mathcal{K}_j} Q(d) \quad \textit{subject to} \quad \|d\| \leq \Delta, \quad \textit{where} \quad Q(d) = \frac{1}{2} d^T B d + g^T d,$$

with corresponding Lagrange multipliers λ_j, $j \in \{1, \ldots, n\}$. If $1 \leq k \leq l \leq n$, then

$$\lambda_k \leq \lambda_l.$$

Proof. Vector d_j is a minimizer of the j-th trust-region subproblem if and only if $\|d_j\| = \|Z_j \tilde{d}_j\| \leq \Delta$, where $Z_j^T (B + \lambda_j I) Z_j \tilde{d}_j = -Z_j^T g$, $Z_j^T (B + \lambda_j I) Z_j \succeq 0$, $\lambda_j \geq 0$ and $\lambda_j(\Delta - \|d_j\|) = 0$, see (11). This minimizer is unconstrained (i.e. the same result is obtained without assuming any trust-region constraint) if and only if $\lambda_j = 0$. If $\lambda_l = 0$, which means that $\|d_l\|$ is the unconstrained minimizer, Lemma 1 implies that $\|d_k\| \leq \|d_l\| \leq \Delta$ for unconstrained minimizer $\|d_k\|$, so $\lambda_k = 0$. If $\lambda_l > 0$ and $\lambda_k = 0$, there is nothing to prove. Let's now suppose that $\lambda_l > 0$ and $\lambda_k > 0$, which means that $\|d_l\| = \|d_k\| = \Delta$. First, assume that $Z_k^T (B + \lambda_k I) Z_k$ is singular and $\lambda_l < \lambda_k$. Then there exists $v \in \mathcal{K}_k$ such that $v^T (B + \lambda_l I) v < 0$ and, since $\mathcal{K}_k \subset \mathcal{K}_l$, $Z_l^T (B + \lambda_l I) Z_l \succeq 0$ cannot hold. This contradiction proves that $\lambda_l \geq \lambda_k$. Assume now that $Z_k^T (B + \lambda_k I) Z_k \succ 0$ and $Z_l^T (B + \lambda_l I) Z_l \succ 0$. As $\mathcal{K}_k(\lambda_k) = \mathcal{K}_k$ by Lemma 2, vector d_k is a solution of unconstrained minimization problem

$$d_k = \arg\min_{d \in \mathcal{K}_k} Q_{\lambda_k}(d), \quad \text{where} \quad Q_\lambda(d) = \frac{1}{2} d^T (B + \lambda I) d + g^T d.$$

Assume that $\lambda_k > \lambda_l$, which implies that $Z_l^T (B + \lambda_k I) Z_l \succ 0$. Let

$$d_l(\lambda_k) = \arg\min_{d \in \mathcal{K}_l} Q_{\lambda_k}(d).$$

Then $\|d_l(\lambda_k)\| \geq \|d_k\| = \Delta$ follows from Lemma 1. Since

$$d_l = \arg\min_{d \in \mathcal{K}_l} Q_{\lambda_l}(d)$$

and $\|d_l\| = \Delta \leq \|d_l(\lambda_k)\|$, Lemma 4 implies that $\lambda_k \leq \lambda_l$ which is a contradiction. Thus $\lambda_k \leq \lambda_l$ has to hold. Finally, assume that $Z_l^T(B+\lambda_l I)Z_l$ is singular. In this case we have $\|d_l(\lambda_l+\varepsilon)\| \leq \Delta$ for arbitrary $\varepsilon > 0$. Since $Z_l^T(B+(\lambda_l+\varepsilon)I)Z_l$ is positive definite, also $Z_k^T(B+(\lambda_l+\varepsilon)I)Z_k$ is positive definite and $\|d_k(\lambda_l+\varepsilon)\| \leq \|d_l(\lambda_l+\varepsilon)\| \leq \Delta$ by Lemma 1. Since $\|d_k\| = \Delta$, Lemma 4 implies that $\lambda_k \leq \lambda_l + \varepsilon$ and, since ε is arbitrary, $\lambda_k \leq \lambda_l$. $\square$

Now we return to subproblem (12). If we set $\tilde{\lambda} = \lambda_k$ for some $k \leq n$, then Theorem 2 implies that $0 \leq \tilde{\lambda} = \lambda_k \leq \lambda_n = \lambda$. As a consequence of this inequality, one has that $\lambda = 0$ implies $\tilde{\lambda} = 0$ so that $\|d\| < \Delta$ implies $\tilde{\lambda} = 0$. Thus the shifted Steihaug-Toint method reduces to the standard one in this case. At the same time, if B is positive definite and $\tilde{\lambda} > 0$, then one has $\Delta \leq \|(B+\tilde{\lambda}I)^{-1}g\| < \|B^{-1}g\|$ by Lemma 4. Thus the unconstrained minimizer of (12) is closer to the trust-region boundary than the unconstrained minimizer of (10) and we can expect that $d(\tilde{\lambda})$ is closer to the optimal locally constrained step than d. Finally, if $\tilde{\lambda} > 0$, then matrix $B+\tilde{\lambda}I$ is better conditioned than B and we can expect that the shifted Steihaug-Toint method will converge more rapidly than the original one. The shifted Steihaug-Toint method consists of the three major steps.

Step 1: Carry out $k \ll n$ steps of the unpreconditioned Lanczos method (described e.g. in [3]) to obtain tridiagonal matrix $T = T_k = Z_k^T B Z_k$.

Step 2: Solve subproblem

$$\text{minimize} \quad (1/2)\tilde{d}^T T \tilde{d} + \|g\| e_1^T \tilde{d} \quad \text{subject to} \quad \|\tilde{d}\| \leq \Delta, \tag{15}$$

using the method of Moré and Sorensen [5], to obtain Lagrange multiplier $\tilde{\lambda}$.

Step 3: Apply the (preconditioned) Steihaug-Toint method to subproblem (12) to obtain direction vector $d = d(\tilde{\lambda})$.

Now we show that the shifted Steihaug-Toint method is globally convergent. Since conditions (2) and (3) are satisfied automatically, it suffices to prove inequality (4) and Theorem 1 can be used (see Corollary 2).

Theorem 3. *Let $d \in \mathcal{R}^n$ be a direction vector obtained by the shifted Steihaug-Toint method with preconditioner C. Then (4) holds with $\underline{\sigma} = 1/(8\kappa(C))$, where $\kappa(C)$ is the spectral condition number of preconditioner C.*

Proof. (a) First, consider the CG method with preconditioner C (symmetric and positive definite) applied to subproblem (12). This method is equivalent to the (unpreconditioned) CG method applied to quadratic function $\hat{Q}(\hat{d}) = (1/2)\hat{d}^T\hat{B}\hat{d} + \hat{g}^T\hat{d}$, where $\hat{d} = C^{1/2}d$, $\hat{g} = C^{-1/2}g$ and $\hat{B} = C^{-1/2}(B+\tilde{\lambda}I)C^{-1/2}$. If at least one CG step is performed, then

$$-\tilde{Q}(d) = -\hat{Q}(\hat{d}) \geq \frac{\|\hat{g}\|^2}{2\|\hat{B}\|} = \frac{g^T C^{-1} g}{2\|C^{-1/2}(B+\tilde{\lambda}I)C^{-1/2}\|} \geq \frac{\|g\|^2}{2\kappa(C)\|B+\tilde{\lambda}I\|}$$

(the first inequality is proved in [8]). If the first CG step lies outside the trust-region, then

$$d_1 = C^{-1/2}\hat{d}_1 = -\frac{\hat{g}^T\hat{g}}{\hat{g}^T\hat{B}\hat{g}} C^{-1/2}\hat{g} = -\frac{g^T C^{-1} g}{g^T C^{-1}(B+\tilde{\lambda}I)C^{-1} g} C^{-1} g$$

implies that

$$\frac{g^T C^{-1} g \sqrt{g^T C^{-2} g}}{g^T C^{-1}(B+\tilde{\lambda}I)C^{-1} g} \geq \Delta \quad \Rightarrow \quad \frac{g^T C^{-1}(B+\tilde{\lambda}I)C^{-1} g}{\sqrt{g^T C^{-2} g}} \Delta \leq g^T C^{-1} g.$$

In this case, $d = (\Delta/||d_1||)d_1 = -(\Delta/\sqrt{g^T C^{-2} g})C^{-1}g$ and we can write

$$-\tilde{Q}(d) = \frac{g^T C^{-1} g}{\sqrt{g^T C^{-2} g}}\Delta - \frac{1}{2}\frac{g^T C^{-1}(B+\tilde{\lambda}I)C^{-1}g}{g^T C^{-2} g}\Delta^2 \geq \frac{1}{2}\frac{g^T C^{-1} g}{\sqrt{g^T C^{-2} g}}\Delta \geq \frac{||g||}{2\kappa(C)}\Delta.$$

Using both inequalities above we obtain

$$-\tilde{Q}(d) \geq \frac{||g||}{2\kappa(C)} \min\left(\Delta, \frac{||g||}{||B+\tilde{\lambda}I||}\right).$$

(b) Since $Z_k^T Z_k = I$ implies

$$\max_{||\tilde{v}||=1} \tilde{v}^T T \tilde{v} = \max_{||\tilde{v}||=1} \tilde{v}^T Z_k^T B Z_k \tilde{v} \leq \max_{||v||=1} v^T B v$$

($\tilde{v} \in \mathcal{R}^k$ and $v \in \mathcal{R}^n$), we can write $||T|| \leq ||B||$. If $\tilde{\lambda} > 0$, then $||\tilde{d}(\tilde{\lambda})|| = \Delta$, where $(T+\tilde{\lambda}I)\tilde{d}(\tilde{\lambda}) = -||g||e_1$ with $||e_1|| = 1$ (see (15) and (11)). Thus

$$||g||^2 = \tilde{d}(\tilde{\lambda})^T (T+\tilde{\lambda}I)^2 \tilde{d}(\tilde{\lambda}) \geq \Delta^2 \min_{||\tilde{d}||=1} \tilde{d}^T (T+\tilde{\lambda}I)^2 \tilde{d} = \Delta^2(\lambda_1+\tilde{\lambda})^2,$$

where λ_1 is the smallest eigenvalue of T. Since $\lambda_1 \geq -||T||$, we can substitute it into the previous inequality to obtain

$$\tilde{\lambda} \leq \frac{1}{\Delta}||g|| + ||T|| \leq \frac{1}{\Delta}||g|| + ||B||.$$

Thus

$$\frac{||B+\tilde{\lambda}I||}{||g||} \leq \frac{2||B||}{||g||} + \frac{1}{\Delta} \leq 2\max\left(\frac{2||B||}{||g||}, \frac{1}{\Delta}\right) \quad \Rightarrow \quad \frac{||g||}{||B+\tilde{\lambda}I||} \geq \frac{1}{2}\min\left(\frac{||g||}{2||B||}, \Delta\right).$$

Using (a) and inequality $\tilde{Q}(d) = Q(d) + \tilde{\lambda}\Delta^2/2 \geq Q(d)$, we can write

$$\begin{aligned} -Q(d) &\geq -\tilde{Q}(d) \geq \frac{1}{2\kappa(C)}||g|| \min\left(\Delta, \frac{||g||}{||B+\tilde{\lambda}I||}\right) \\ &\geq \frac{1}{2\kappa(C)}||g|| \min\left(\Delta, \frac{1}{2}\min\left(\frac{||g||}{2||B||}, \Delta\right)\right) \geq \frac{1}{8\kappa(C)}||g|| \min\left(\Delta, \frac{||g||}{||B||}\right) \end{aligned}$$

and (4) holds with $\underline{\sigma} = 1/(8\kappa(C))$. □

Corollary 2. *If there exist constants $\overline{B}$ and $\overline{C}$ such that matrices B_i and preconditioners C_i satisfy conditions $||B_i|| \leq \overline{B}$, $\kappa(C_i) \leq \overline{C}$ $\forall i \in \mathcal{N}$, then the shifted Steihaug-Toint method is globally convergent in the sense of Theorem 1.*

3 Computational experiments

Now we present a numerical comparison of seven methods for computing direction vectors satisfying conditions (2)-(4):

- MS - the method of Moré and Sorensen [5] for computing the optimal locally constrained step.

- DL - the dog-leg strategy of Powell [6] or Dennis and Mei [2].
- MDL - the multiple dog-leg strategy ($k = 5$) mentioned in [8].
- ST - the basic (unpreconditioned) Steihaug [8] and Toint [9] method.
- GLRT - the method of Gould, Lucidi, Roma and Toint [3] which combines CG method with the Lanczos process to give a good approximation of the optimal locally constrained step.
- PST - Preconditioned Steihaug-Toint method (with the incomplete Choleski preconditioner).
- PSST - Preconditioned shifted Steihaug-Toint method (with the incomplete Choleski preconditioner), $k = 5$.

These algorithms were used for solving trust-region subproblems arising in a realization of a discrete Newton method. They were tested by using a set of 22 sparse least-squares test problems with 1000 and 5000 variables (subroutine `TEST14` [4], which can be found on the page `www.cs.cas.cz/~luksan/test.html`). The results are given in Table 1, where `NIT` is the total number of iterations, `NFV` is the total number of function evaluations, `NFG` is the total number of gradient evaluations, `NCG` is the total number of the CG iterations and `Time` is the total computational time (in seconds).

N	Method	NIT	NFV	NFG	NCG	Time
1000	MS	1918	1955	8797	-	4.65
	DL	2515	2716	11859	-	4.42
	MDL	2292	2456	10673	12203	4.61
	ST	3329	3784	16456	53573	8.20
	GLRT	3107	3444	15306	55632	8.53
	PST	2631	2823	13019	910	5.14
	PSST	1999	2046	9201	1161	4.25
5000	MS	8391	8566	35824	-	122.44
	DL	9657	10133	42425	-	115.77
	MDL	8938	9276	39032	47236	122.84
	ST	16894	19163	83933	358111	364:42
	GLRT	14679	16383	71483	366695	401.45
	PST	10600	11271	50365	3767	145.42
	PSST	8347	8454	35939	4329	108.87

Table 1 : Comparison of methods using TEST14.

For a better comparison of methods PST, PSST, DL and MS, we have performed additional tests with problems from the widely used CUTE collection [1]. Table 2 contains a list of these problems together with their dimensions and results obtained. Values `NIT`, `NFV`, `NFG`, `NCG` and `Time` have the same meaning as in the previous table.

Table 2 implies several conclusions. If problems do not have sparse Hessian matrices, then direct methods DL and MS can be much worse than iterative methods PST and PSST as is demonstrated on problems MSQRTALS, NONCVXU2, NONCVXUN and SPARSINE. On the other hand, direct methods can be more efficient for ill-conditioned but reasonably sparse problems, e.g., CHAINWOO and SBRYBND. Comparing PST and PSST, we can see that PSST is usually slightly worse than PST, measured by the computational time, since it uses additional operations for determining the Lanczos matrix T and computing parameter $\tilde{\lambda}$. Nevertheless, if the problems are difficult as BROWNAL, CHAINWOO, FMINSURF, MSQRTALS and NONCVXUN, then PSST is much better than PST. Thus the total computational time can be lower for PSST as in Table 1.

Method		PST					PSST					DL					MS				
Problem	N	NIT	NFV	NFG	NCG	Time	NIT	NFV	NFG	NCG	Time	NIT	NFV	NFG	NCG	Time	NIT	NFV	NFG	NCG	Time
ARWHEAD	5000	7	18	21	5	0.72	6	7	21	6	0.67	6	17	18	5	0.69	6	17	18	22	0.70
BDQRTIC	5000	14	25	126	9	2.63	14	15	135	13	2.72	13	14	126	13	2.63	14	15	135	21	2.89
BROWNAL	500	8	9	4509	6	183.61	4	5	2505	4	156.28	4	5	2505	3	155.53	5	6	3006	8	160.76
BROYDN7D	2000	49	56	450	193	1.94	55	66	504	169	2.25	80	105	729	60	3.19	36	42	333	150	1.83
BRYBND	5000	11	13	168	8	2.16	13	16	196	16	2.48	16	19	238	11	2.89	15	18	224	38	3.22
CHAINWOO	1000	963	1246	4815	8653	6.99	747	952	3740	5749	5.92	386	478	1935	374	2.80	44	53	225	115	0.36
COSINE	5000	5	7	24	4	0.31	6	8	28	9	0.39	5	7	24	4	0.36	15	16	64	37	0.67
CRAGGLVY	5000	18	19	76	17	1.04	18	19	76	18	1.11	18	19	76	17	1.06	17	18	72	150	1.06
CURLY10	1000	21	33	462	32	0.49	19	21	440	23	0.45	18	20	418	15	0.44	21	23	484	77	0.72
CURLY20	1000	18	21	798	65	1.41	18	20	798	51	1.51	15	18	672	13	1.27	15	17	672	60	1.79
CURLY30	1000	19	22	1240	80	3.08	16	17	1054	71	2.77	12	13	806	11	2.14	18	20	1178	72	4.09
DIXMAANA	3000	6	7	56	4	0.34	5	6	48	5	0.33	5	6	48	4	0.34	5	6	48	8	0.34
DIXMAANB	3000	6	7	56	5	0.34	7	8	64	7	0.38	12	16	104	8	0.54	11	14	96	38	0.62
DIXMAANC	3000	7	8	64	7	0.37	7	8	64	7	0.39	10	13	88	6	0.47	9	11	80	24	0.49
DIXMAAND	3000	8	9	72	7	0.39	8	9	72	8	0.41	22	28	184	16	0.82	11	13	96	34	0.59
DIXMAANE	3000	8	9	72	11	0.41	7	8	64	8	0.39	21	25	176	15	0.76	10	12	88	32	0.58
DIXMAANF	3000	14	16	120	28	0.55	15	17	128	30	0.61	36	46	296	21	1.22	24	30	200	96	1.13
DIXMAANG	3000	14	15	120	32	0.56	15	17	128	33	0.59	47	60	384	28	1.51	20	25	168	83	0.95
DIXMAANH	3000	16	19	136	104	0.64	16	19	136	101	0.69	72	91	584	56	2.20	24	30	200	97	1.13
DIXMAANI	3000	10	12	88	11	0.47	9	10	80	9	0.44	19	24	160	15	0.77	10	12	88	23	0.55
DIXMAANJ	3000	19	24	160	1825	1.18	21	26	176	669	0.97	69	90	560	49	2.12	34	42	280	150	1.53
DIXMAANK	3000	21	26	176	1145	1.05	23	28	192	1020	1.11	105	130	848	97	3.24	22	28	184	87	1.05
DIXMAANL	3000	25	30	208	556	1.03	23	28	192	1428	1.27	172	221	1384	164	5.26	25	31	208	110	1.17
DQRTIC	5000	33	34	68	33	0.30	34	35	70	33	0.37	34	35	70	33	0.39	34	35	70	35	0.33
EDENSCH	5000	12	13	52	11	0.70	11	12	48	11	0.70	32	40	132	19	1.42	12	13	52	12	0.72
EG2	1000	3	4	12	3	0.02	6	7	21	6	0.06	3	4	12	3	0.06	13	25	39	28	0.06
ENGVAL1	5000	8	9	36	7	0.50	8	9	36	8	0.55	8	9	36	7	0.53	8	9	36	8	0.50
EXTROSNB	1000	4756	5036	19028	4228	18.68	4735	4977	18944	4208	20.42	4746	5002	18988	4745	19.19	4597	4656	18392	5621	18.45
FLETCBV2	1000	7	8	32	0	0.08	7	8	32	6	0.09	8	9	36	8	0.09	7	8	32	14	0.06
FLETCHCR	1000	1436	1440	5748	723	5.89	1433	1435	5736	726	6.39	1437	1441	5752	1436	6.34	1396	1402	5588	2095	6.12

Table 2a : Comparison of methods using the CUTE problems.

Method		PST					PSST					DL					MS				
Problem	N	NIT	NFV	NFG	NCG	Time	NIT	NFV	NFG	NCG	Time	NIT	NFV	NFG	NCG	Time	NIT	NFV	NFG	NCG	Time
FMINSRF2	1024	21	28	220	35	0.30	54	60	550	172	0.72	87	88	880	86	1.27	45	52	460	187	1.17
FMINSURF	484	142	158	69355	326	119.95	46	49	22795	89	62.02	25	32	12610	24	40.24	42	45	20855	161	69.27
FREUROTH	5000	8	21	32	6	0.61	8	12	36	10	0.66	7	10	32	5	0.59	7	10	32	11	0.58
GENHUPPS	1000	8121	8637	32488	13073	57.03	7466	7775	29868	23576	56.80	2087	2451	8352	2014	15.54	NFG > 80000				
GENROSE	1000	755	892	3024	3708	3.39	722	864	2892	3419	3.55	764	942	3060	759	3.44	709	875	2840	2576	3.95
LIARWHD	1000	15	17	48	22	0.67	15	17	48	27	0.74	13	14	42	12	0.65	13	14	42	14	0.63
MOREBV	5000	17	18	108	1	0.89	17	18	108	4	0.92	17	18	108	16	0.95	4	5	30	30	0.63
MSQRTALS	529	33	41	18020	1501	104.30	30	36	16430	990	97.56	NFG > 40000					53	69	28620	429	246.61
NCB20	510	56	62	2565	456	6.53	47	53	2160	541	5.48	231	296	10440	221	26.39	72	82	3285	431	9.61
NCB20B	1010	55	66	2464	519	14.49	48	55	2156	420	12.62	287	384	12672	260	74.28	44	54	1980	270	13.91
NONCVXU2	1000	376	396	4524	1403	4.38	263	304	3168	1943	3.54	825	1226	9912	463	120.81	298	344	3588	1297	303.50
NONCVXUN	1000	1168	1185	14028	791145	104.05	611	644	7344	406737	53.99	NFG > 80000					NFG > 80000				
NONDIA	5000	6	7	21	7	0.47	7	8	24	9	0.52	5	6	18	4	0.45	5	6	18	5	0.44
NONDQUAR	5000	26	27	135	53	1.19	24	26	125	141	1.27	21	22	110	20	1.13	20	22	105	127	1.56
PENALTY1	500	41	44	21042	36	60.30	41	44	21042	39	63.05	41	44	21042	37	49.00	40	44	20541	54	52.33
POWELLSG	5000	17	18	72	40	0.45	17	18	72	41	0.50	17	18	72	16	0.45	17	18	72	21	0.45
POWER	500	28	29	14529	27	48.03	28	29	14529	27	47.84	28	29	14529	27	37.78	28	29	14529	30	40.31
QUARTC	5000	231	232	464	42	1.77	224	225	450	99	2.29	231	232	464	21	1.64	231	232	464	231	1.88
SBRYBND	5000	NFG > 40000					NFG > 40000					38	42	546	37	5.97	27	29	392	93	5.69
SCHMVETT	5000	3	4	24	2	0.56	3	4	24	3	0.59	5	6	36	5	0.74	3	4	24	3	0.58
SCOSINE	5000	36	44	148	11	1.17	36	44	148	11	1.29	31	53	128	6	1.09	8	22	32	109	0.79
SINQUAD	5000	223	242	896	495	9.86	225	236	904	565	10.53	49	60	200	49	2.97	224	250	900	549	10.45
SPARSINE	1000	13	15	924	1369	2.27	14	16	990	1614	2.59	NFG > 80000					24	33	1650	263	291.64
SPARSQUR	1000	19	20	1320	21	1.54	19	20	1320	22	1.57	18	19	1254	17	20.09	19	20	1320	20	23.30
SPMSRTLS	4999	14	20	135	72	1.67	16	21	153	41	1.83	35	44	324	24	3.25	17	22	162	63	2.30
SROSENBR	5000	9	11	30	7	0.27	6	7	21	6	0.23	6	7	21	5	0.24	6	7	21	8	0.24
TOINTGSS	5000	2	3	18	1	0.38	9	13	60	24	0.74	3	4	24	0	0.44	1	2	12	1	0.33
TQUARTIC	5000	20	21	63	12	0.74	17	18	54	26	0.74	16	17	51	15	0.68	16	17	51	36	0.70
VAREIGVL	500	14	15	7515	13	49.70	13	14	7014	13	49.28	15	18	8016	14	46.25	14	15	7515	14	45.53
WOODS	4000	42	49	172	41	0.81	41	48	168	51	0.91	40	47	164	39	0.81	40	46	164	80	0.88

Table 2b : Comparison of methods using the CUTE problems.

Reference

[1] I.Bongartz, A.R.Conn, N.Gould, P.L.Toint: CUTE: constrained and unconstrained testing environment. ACM Transactions on Mathematical Software 21, 123-160, 1995.

[2] J.E.Dennis, H.H.W.Mei: An unconstrained optimization algorithm which uses function and gradient values. Report No. TR 75-246, 1975.

[3] N.I.M.Gould, S.Lucidi, M.Roma, P.L.Toint: Solving the trust-region subproblem using the Lanczos method. Report No. RAL-TR-97-028, 1997.

[4] L.Lukšan, J.Vlček: Sparse and partially separable test problems for unconstrained and equality constrained optimization. Report No. V767-98, Institute of Computer Science AVČR, 1998.

[5] J.J.Moré, D.C.Sorensen: Computing a trust region step. SIAM Journal on Scientific and Statistical Computations 4 (1983) 553-572.

[6] M.J.D.Powell: A new algorithm for unconstrained optimization. In: "Nonlinear Programming" (J.B.Rosen O.L.Mangasarian, K.Ritter, eds.) Academic Press, London 1970.

[7] M.J.D.Powell: On the global convergence of trust region algorithms for unconstrained optimization. Mathematical Programming 29 (1984) 297-303.

[8] T.Steihaug: The conjugate gradient method and trust regions in large-scale optimization. SIAM Journal on Numerical Analysis 20 (1983) 626-637.

[9] P.L.Toint: Towards an efficient sparsity exploiting Newton method for minimization. In: Sparse Matrices and Their Uses (I.S.Duff, ed.), Academic Press, London 1981, 57-88.