Efficient tridiagonal preconditioner for the matrix-free truncated Newton method

Lukšan, Ladislav
2013

Institute of Computer Science
Academy of Sciences of the Czech Republic

Efficient tridiagonal preconditioner for the matrix-free truncated Newton method

Ladislav Lukšan, Jan Vlček

Technical report No. 1177

January 2013

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

Efficient tridiagonal preconditioner for the matrix-free truncated Newton method

Ladislav Lukšan, Jan Vlček [1]

Technical report No. 1177

January 2013

Abstract:

In this paper, we study an efficient tridiagonal preconditioner, based on numerical differentiation, applied to the matrix-free truncated Newton method for unconstrained optimization. It is proved that this preconditioner is positive definite for many practical problems. The efficiency of the resulting matrix-free truncated Newton method is demonstrated by results of extensive numerical experiments.

Keywords:

[1]This work was supported by the Institute of Computer Science of the AS CR (RVO:67985807)

1 Introduction

We consider the unconstrained minimization problem

$$x^* = \arg\min_{x \in R^n} F(x),$$

where function $F : \mathcal{D}(F) \subset R^n \to R$ is twice continuously differentiable and n is large. We use the notation

$$g(x) = \nabla F(x), \qquad G(x) = \nabla^2 F(x)$$

and the assumption that $\|G(x)\| \leq \overline{G}$, $\forall x \in \mathcal{D}(F)$. Numerical methods for unconstrained minimization are usually iterative and their iteration step has the form

$$x_{k+1} = x_k + \alpha_k d_k, \quad k \in N,$$

where d_k is a direction vector and α_k is a step-length. In this paper, we will deal with the Newton method, which uses the quadratic model

$$F(x_k + d) \approx Q(x_k + d) = F(x_k) + g^T(x_k)d + \frac{1}{2} d^T G(x_k) d \tag{1}$$

for direction determination in such a way that

$$d_k = \arg\min_{d \in \mathcal{M}_k} Q(x_k + d). \tag{2}$$

There are two basic possibilities for direction determination: the line-search method, where $\mathcal{M}_k = R^n$, and the trust-region method, where $\mathcal{M}_k = \{d \in R^n : \|d\| \leq \Delta_k\}$ (here $\Delta_k > 0$ is the trust region radius). Properties of line search and trust region methods together with comments concerning their implementation are exhaustively introduced in [3], [19], so no more details are given here.

In this paper, we assume that neither matrix $G_k = G(x_k)$ nor its sparsity pattern are explicitly known. In this case, direct methods based on Gaussian elimination cannot be used, so it is necessary to compute the direction vector (2) iteratively. There are many various iterative methods making use of a symmetry of the Hessian matrix, see [23]. Some of them, e.g. [7], [8], [21] allow us to consider indefinite Hessian matrices. Even if these methods are of theoretical interest and lead to nontraditional preconditioners, see [9] and [10], we confine our attention to modifications of the conjugate gradient method [24], [25], [26], which are simple and very efficient (also in the indefinite case). We studied and tested both the line search and the trust region approaches, but the second approach did not give significantly better results than the first one. Therefore, we restrict our attention to the line search implementation of the truncated Newton method.

Since matrix $G(x)$ is not given explicitly, we use numerical differentiation instead of matrix multiplication. Thus the product $G(x)p$ is replaced by the difference

$$G(x)p \approx \frac{g(x + \delta p) - g(x)}{\delta} \tag{3}$$

where $\delta = \varepsilon/\|p\|$ (usually $\varepsilon \approx \sqrt{\varepsilon_M}$ where ε_M is the machine precision). The resulting method is called the truncated Newton method. This method has been theoretically studied in many papers, see [4], [5], [17], [20]. The following theorem, which easily follows from the mean value theorem, confirms the choice (3).

Theorem 1 *Let function $F : R^n \to R$ have Lipschitz continuous second order derivatives (with the Lipschitz constant $\overline{L}$). Let $q = G(x)p$ and*

$$\tilde{q} = \frac{g(x + \delta p) - g(x)}{\delta}, \quad \delta = \frac{\varepsilon}{\|p\|}.$$

Then it holds

$$\|\tilde{q} - q\| \leq \frac{1}{2}\varepsilon\overline{L}\|p\|.$$

To make the subsequent investigations clear, we briefly describe the preconditioned conjugate gradient subalgorithm proposed in [24] where matrix multiplications are replaced by gradient differences (the outer index k is for the sake of simplicity omitted).
Truncated Newton PCG subalgorithm:

$d_1 = 0, \quad g_1 = g, \quad h_1 = C^{-1}g_1, \quad \rho_1 = g_1^T h_1, \quad p_1 = -h_1.$
Do $i = 1$ **to** $n + 3$
$\delta_i = \varepsilon/\|p_i\|, \quad \tilde{q}_i = (g(x + \delta p_i) - g(x))/\delta_i, \quad \sigma_i = p_i^T \tilde{q}_i.$
If $\sigma_i < \varepsilon\|p_i\|^2$ **then** $d = d_i,$ stop.
$\alpha_i = \rho_i/\sigma_i, \quad d_{i+1} = d_i + \alpha_i p_i, \quad g_{i+1} = g_i + \alpha_i \tilde{q}_i,$
$h_{i+1} = C^{-1}g_{i+1}, \quad \rho_{i+1} = g_{i+1}^T h_{i+1}.$
If $\|g_{i+1}\| \leq \omega\|g_1\|$ **or** $i = m$ **then** $d = d_i,$ stop.
$\beta_i = \rho_{i+1}/\rho_i, \quad p_{i+1} = -h_{i+1} + \beta_i p_i.$
End do

A disadvantage of the truncated Newton PCG subalgorithm with $C = I$ (unpreconditioned) consists in the fact that it requires a large number of inner iterations (i.e. a large number of gradient evaluations) if matrix $G = G(x)$ is ill-conditioned. Thus a suitable preconditioner should be used. Unfortunately, the sparsity pattern of G is not known, so the standard preconditioning methods requiring the knowledge of the sparsity pattern (e.g. methods based on the incomplete Choleski decomposition) cannot be chosen.

There are various ways for building positive definite preconditioners, which can be utilized in the truncated Newton PCG subalgorithm:

- Preconditioners based on the limited memory BFGS updates. This very straightforward approach is studied in [12] and [16].
- Band preconditioners obtained by the standard BFGS method equivalent to the preconditioned conjugate gradient method. This approach is described in [18], where it is used for building diagonal preconditioners. More general band preconditioners of this type are studied in [14].

- Band preconditioners obtained by numerical differentiation. This approach is used in [22] for building diagonal preconditioners. More general band preconditioners of this type are studied in [14].
- Preconditioners determined by the Lanczos method equivalent to the conjugate gradient method. This approach is studied in [9], [10] and [17].

In this paper, we propose new results concerning tridiagonal preconditioners obtained by numerical differentiation and show that they are very efficient in connection with the truncated Newton method. This efficiency can be observed from tables and figures introduced in Section 3, where the comparison of our two implementations of the tridiagonally preconditioned truncated Newton method with the unpreconditioned method and the method preconditioned by the limited memory BFGS updates is given.

2 Tridiagonal preconditioners based on the numerical differentiation

If the Hessian matrix is tridiagonal, its elements can be simply approximated by numerical differentiation. If the Hessian matrix is not tridiagonal, we can use this process to determine a suitable tridiagonal preconditioner. Numerical differentiation is performed only once at the beginning of the outer step of the Newton method.

In order to determine all elements of the tridiagonal Hessian matrix, it suffices to use two gradient differences $g(x+\varepsilon v_1)-g(x)$ and $g(x+\varepsilon v_2)-g(x)$, where $v_1 = [\delta_1, 0, \delta_3, 0, \delta_5, 0, \ldots]$, $v_2 = [0, \delta_2, 0, \delta_4, 0, \delta_6, \ldots]$, and $\varepsilon > 0$ (usually $\varepsilon \approx \sqrt{\varepsilon_M}$), which means to compute two extra gradients during each outer step of the Newton method. The differences δ_i, $1 \le i \le n$, can be chosen by two different ways:

(1) We set $\delta_i = \delta$, $1 \le i \le n$, where $\delta > 0$ (usually $\delta \approx \sqrt{2/n}$).

(2) We set $\delta_i = \max(|x_i|, 1)$, $1 \le i \le n$.

Theorem 2 *Let the Hessian matrix of function $F : R^n \to R$ be tridiagonal of the form*

$$T = \begin{bmatrix} \alpha_1 & \beta_1 & \ldots & 0 & 0 \\ \beta_1 & \alpha_2 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \alpha_{n-1} & \beta_{n-1} \\ 0 & 0 & \ldots & \beta_{n-1} & \alpha_n \end{bmatrix}. \tag{4}$$

Set $v_1 = [\delta_1, 0, \delta_3, 0, \delta_5, 0, \ldots]$, $v_2 = [0, \delta_2, 0, \delta_4, 0, \delta_6, \ldots]$, where $\delta_i > 0$, $1 \le i \le n$. Then for $2 \le i \le n-1$ one has

$$\alpha_1 = \lim_{\varepsilon \to 0} \frac{g_1(x+\varepsilon v_1) - g_1(x)}{\varepsilon \delta_1}, \quad \beta_1 = \lim_{\varepsilon \to 0} \frac{g_1(x+\varepsilon v_2) - g_1(x)}{\varepsilon \delta_2},$$
$$\alpha_i = \lim_{\varepsilon \to 0} \frac{g_i(x+\varepsilon v_1) - g_i(x)}{\varepsilon \delta_i}, \quad \beta_i = \lim_{\varepsilon \to 0} \frac{g_i(x+\varepsilon v_2) - g_i(x)}{\varepsilon \delta_{i+1}} - \beta_{i-1}\frac{\delta_{i-1}}{\delta_{i+1}}, \quad \mathrm{mod}(i,2) = 1,$$

$$\alpha_i = \lim_{\varepsilon\to 0} \frac{g_i(x+\varepsilon v_2) - g_i(x)}{\varepsilon\delta_i}, \quad \beta_i = \lim_{\varepsilon\to 0} \frac{g_i(x+\varepsilon v_1) - g_i(x)}{\varepsilon\delta_{i+1}} - \beta_{i-1}\frac{\delta_{i-1}}{\delta_{i+1}}, \quad \mathrm{mod}(i,2) = 0,$$
$$\alpha_n = \lim_{\varepsilon\to 0} \frac{g_n(x+\varepsilon v_1) - g_n(x)}{\varepsilon\delta_n}, \quad \mathrm{mod}(n,2) = 1,$$
$$\alpha_n = \lim_{\varepsilon\to 0} \frac{g_n(x+\varepsilon v_2) - g_n(x)}{\varepsilon\delta_n}, \quad \mathrm{mod}(n,2) = 0.$$

Proof Theorem 1 implies that $g(x+\varepsilon v_1) - g(x) = \varepsilon G(x)v_1 + o(\varepsilon)$, $g(x+\varepsilon v_2) - g(x) = \varepsilon G(x)v_2 + o(\varepsilon)$, so after substituting $G(x) = T$, where T is a tridiagonal matrix of the form (4), and rearranging individual elements we obtain

$$\frac{g_1(x+\varepsilon v_1) - g_1(x)}{\varepsilon\delta_1} = \alpha_1 + o(1), \qquad \frac{g_1(x+\varepsilon v_2) - g_1(x)}{\varepsilon\delta_2} = \beta_1 + o(1),$$
$$\frac{g_i(x+\varepsilon v_1) - g_i(x)}{\varepsilon\delta_i} = \alpha_i + o(1), \qquad \frac{g_i(x+\varepsilon v_2) - g_i(x)}{\varepsilon\delta_i} = \beta_i + \beta_{i-1}\frac{\delta_{i-1}}{\delta_{i+1}} + o(1), \qquad \mathrm{mod}(i,2) = 1,$$
$$\frac{g_i(x+\varepsilon v_2) - g_i(x)}{\varepsilon\delta_i} = \alpha_i + o(1), \qquad \frac{g_i(x+\varepsilon v_1) - g_i(x)}{\varepsilon\delta_i} = \beta_i + \beta_{i-1}\frac{\delta_{i-1}}{\delta_{i+1}} + o(1), \qquad \mathrm{mod}(i,2) = 0,$$
$$\frac{g_n(x+\varepsilon v_1) - g_i(x)}{\varepsilon\delta_i} = \alpha_n + o(1), \qquad \mathrm{mod}(n,2) = 1,$$
$$\frac{g_n(x+\varepsilon v_2) - g_i(x)}{\varepsilon\delta_i} = \alpha_n + o(1), \qquad \mathrm{mod}(n,2) = 0,$$

where $2 \le i \le n-1$. Since ratios $\delta_{i-1}/\delta_{i+1}$, $2 \le i \le n-1$, are indepenent of ε, the theorem is proved. □

Remark 1 Theorem 2 specifies an efficient way for building a tridiagonal preconditioner. We choose fixed numbers ε, δ_i, $1 \le i \le n$, and compute elements of the tridiagonal matrix $C = T(\varepsilon)$ according to formulas mentioned in Theorem 2, where the limits are omitted. Denoting by $T(0)$ the matrix appearing in Theorem 2, we can write $T(0) = \lim_{\varepsilon\downarrow 0} T(\varepsilon)$. Then $G(x)v_1 = T(0)v_1$ and $G(x)v_2 = T(0)v_2$, where $v_1 = [\delta_1, 0, \delta_3, 0, \delta_5, 0, \ldots]^T$ and $v_2 = [0, \delta_2, 0, \delta_4, 0, \delta_6, \ldots]^T$. If $\delta_i = \delta$, $1 \le i \le n$ (all differences are the same), the elements of the matrix $T(0)$ can be expressed in the form

$$\alpha_i = \sum_{\mathrm{mod}(j,2)=1} G_{ij}, \qquad \beta_i + \beta_{i-1} = \sum_{\mathrm{mod}(j,2)=0} G_{ij}, \qquad \mathrm{mod}(i,2) = 1,$$
$$\alpha_i = \sum_{\mathrm{mod}(j,2)=0} G_{ij}, \qquad \beta_i + \beta_{i-1} = \sum_{\mathrm{mod}(j,2)=1} G_{ij}, \qquad \mathrm{mod}(i,2) = 0,$$

where $\beta_0 = \beta_n = 0$.

Tridiagonal matrix $T(\varepsilon)$ obtained by Remark 1 may not be positive definite even if the Hessian matrix $G(x)$ is positive definite and diagonally dominant.

Example 1 Consider the strictly convex quadratic function $F : R^4 \to R$ with the constant Hessian matrix

$$G = \begin{bmatrix} 7 & 0 & -2 & 4 \\ 0 & 7 & 0 & -2 \\ -2 & 0 & 7 & 0 \\ 4 & -2 & 0 & 7 \end{bmatrix}.$$

Setting $v_1 = [\delta, 0, \delta, 0]^T$, $v_2 = [0, \delta, 0, \delta]^T$, we can write

$$\frac{g(x+v_1)-g(x)}{\delta} = G\begin{bmatrix}1\\0\\1\\0\end{bmatrix} = \begin{bmatrix}5\\0\\5\\4\end{bmatrix}, \qquad \frac{g(x+v_2)-g(x)}{\delta} = G\begin{bmatrix}0\\1\\0\\1\end{bmatrix} = \begin{bmatrix}4\\5\\0\\5\end{bmatrix}$$

and using Theorem 2 we obtain

$$T(\varepsilon) = T(0) = \begin{bmatrix}5 & 4 & 0 & 0\\ 4 & 5 & -4 & 0\\ 0 & -4 & 5 & 4\\ 0 & 0 & 4 & 5\end{bmatrix}.$$

This matrix is not positive definite, since determinant

$$\det\begin{bmatrix}5 & 4 & 0\\ 4 & 5 & -4\\ 0 & -4 & 5\end{bmatrix} = 5\,(25-32) = -35$$

of its principal submatrix is negative.

The above example shows, that the diagonal dominance of the Hessian matrix does not suffice for positive definiteness of the tridiagonal matrix obtained by Remark 1. First we show that this condition is sufficient in case the Hessian matrix is pentadiagonal.

Theorem 3 *Let the Hessian matrix $G(x)$ be pentadiagonal and diagonally dominant with positive diagonal elements. Then if $\delta_i = \delta$, $1 \le i \le n$, and the number $\varepsilon > 0$ is sufficiently small, the tridiagonal matrix $T(\varepsilon)$ obtained by Remark 1 is positive definite and diagonally dominant.*

Proof Consider a pentadiagonal Hessian matrix of the form

$$G(x) = \begin{bmatrix}
\tilde\alpha_1 & \tilde\beta_1 & \tilde\gamma_1 & \dots & 0 & 0 & 0\\
\tilde\beta_1 & \tilde\alpha_2 & \tilde\beta_2 & \dots & 0 & 0 & 0\\
\tilde\gamma_1 & \tilde\beta_2 & \tilde\alpha_3 & \dots & 0 & 0 & 0\\
\dots & \dots & \dots & \dots & \dots & \dots & \dots\\
0 & 0 & 0 & \dots & \tilde\alpha_{n-2} & \tilde\beta_{n-2} & \tilde\gamma_{n-2}\\
0 & 0 & 0 & \dots & \tilde\beta_{n-2} & \tilde\alpha_{n-1} & \tilde\beta_{n-1}\\
0 & 0 & 0 & \dots & \tilde\gamma_{n-2} & \tilde\beta_{n-1} & \tilde\alpha_n
\end{bmatrix} \tag{5}$$

and set $\tilde\gamma_{-1} = \tilde\gamma_0 = \tilde\beta_0 = \beta_0 = 0$, $\tilde\gamma_{n-1} = \tilde\gamma_n = \tilde\beta_n = \beta_n = 0$ to simplify the notation. The diagonal dominance of $G(x)$ implies that

$$\tilde\alpha_i > |\tilde\gamma_{i-2}| + |\tilde\beta_{i-1}| + |\tilde\beta_i| + |\tilde\gamma_i|$$

for $1 \le i \le n$. Let $\delta_i = \delta$, for $1 \le i \le n$. Using Theorem 2 (and formulas introduced in Remark 1), the elements of the matrix $T(0)$ can be expressed in the form

$$\alpha_i = \tilde{\gamma}_{i-2} + \tilde{\alpha}_i + \tilde{\gamma}_i, \qquad \beta_{i-1} + \beta_i = \tilde{\beta}_{i-1} + \tilde{\beta}_i \tag{6}$$

for $1 \le i \le n$. Therefore, one has $\beta_i = \tilde{\beta}_i$, which together with (6) gives

$$\alpha_i - |\beta_{i-1}| - |\beta_i| = \tilde{\alpha}_i + \tilde{\gamma}_{i-2} + \tilde{\gamma}_i - |\tilde{\beta}_{i-1}| - |\tilde{\beta}_i| \ge \tilde{\alpha}_i - |\tilde{\gamma}_{i-2}| - |\tilde{\beta}_{i-1}| - |\tilde{\beta}_i| - |\tilde{\gamma}_i| > 0$$

for $1 \le i \le n$. This implies that symmetric tridiagonal matrix $T(0)$ appearing in Theorem 2 has positive elements on the main diagonal and is diagonally dominant. Therefore, it is positive definite (Gershgorin cycles lie in the interior of the right halfplane). Then also the matrix $T(\varepsilon)$ is positive definite for a sufficiently small value $\varepsilon > 0$ (this follows from the continuous dependence of eigenvalues on matrix elements). □

Diagonal dominance of the Hessian matrix is a very strong condition. Nevertheless, the tridiagonal matrix $T(\varepsilon)$ obtained by Remark 1 is positive definite also for other important problems with pentadiagonal Hessian matrices.

Theorem 4 *Let the Hessian matrix $G(x)$ be pentadiagonal and have the form (5) with*

$$\begin{aligned} \tilde{\alpha}_1 \ge \psi_1^2 + 1, \quad \tilde{\alpha}_n \ge \psi_n^2 + 1, \quad \tilde{\alpha}_i \ge \psi_i^2 + 2, \qquad & 2 \le i \le n-1, \\ |\tilde{\beta}_i| \le |\psi_i + \psi_{i+1}|, \qquad & 1 \le i \le n-1, \\ \tilde{\gamma}_i \ge 1, \qquad & 1 \le i \le n-2, \end{aligned} \tag{7}$$

where ψ_i, $1 \le i \le n$, are arbitrary real numbers such that at least one of expressions $\psi_1\psi_2 - 2$, $\psi_i\psi_{i+1} - 4$, $2 \le i \le n-1$, $\psi_{n-1}\psi_n - 2$ is nonzero. Then, if $\delta_i = \delta$, $1 \le i \le n$, and the number $\varepsilon > 0$ is sufficiently small, the matrix $T(\varepsilon)$ obtained by Remark 1 is positive definite.

Proof (a) Let T be a tridiagonal matrix such that

$$2v^T T v = \sum_{i=1}^{n-1} [v_i, v_{i+1}] \begin{bmatrix} \lambda_i & \mu_i \\ \mu_i & \lambda_{i+1} \end{bmatrix} \begin{bmatrix} v_i \\ v_{i+1} \end{bmatrix} \tag{8}$$

for all $v \in R^n$ (for T given by (4), we can set, e.g., $\lambda_1 = 2\alpha_1$, $\lambda_i = \alpha_i$, $2 \le i \le n-1$, $\lambda_n = 2\alpha_n$ and $\mu_i = 2\beta_i$, $1 \le i \le n-1$) We show that if all 2×2 matrices in (8) have positive diagonal elements, are positive semidefinite and at least one of them is positive definite, then T is positive definite. If all 2×2 matrices in (8) are positive semidefinite, one has $v^T T v \ge 0$. Let $v^T T v = 0$, so all terms in (8) are zeroes. If the 2×2 matrix in the i-th term is positive definite, then necessarily $v_i = 0$, $v_{i+1} = 0$. Since the previous and the next terms are zeroes, we can write

$$\begin{aligned} \lambda_{i-1} v_{i-1}^2 + \lambda_i v_i^2 + 2\mu_{i-1} v_{i-1} v_i &= 0, \\ \lambda_{i+1} v_{i+1}^2 + \lambda_{i+2} v_{i+2}^2 + 2\mu_{i+1} v_{i+1} v_{i+2} &= 0, \end{aligned}$$

where $v_i = 0$, $v_{i+1} = 0$ and $\lambda_{i-1} > 0$, $\lambda_{i+2} > 0$. Thus $v_{i-1} = 0$ and $v_{i+2} = 0$. Continuing in this way, we obtain $v_i = 0$ for all $1 \le i \le n$.

(b) If the Hessian matrix $G(x)$ is pentadiagonal of the form (5), we can express the elements of tridiagonal matrix $T(0)$ by the formulas (6) (where $\tilde\gamma_{-1} = \tilde\gamma_0 = \tilde\beta_0 = \beta_0 = 0$ and $\tilde\gamma_{n-1} = \tilde\gamma_n = \tilde\beta_n = \beta_n = 0$). Substitute inequalities (7) into (6), we can write $\alpha_1 \ge \psi_1^2 + 2$, $\alpha_2 \ge \psi_2^2 + 3$, $\alpha_i \ge \psi_i^2 + 4$, $3 \le i \le n-2$, $\alpha_{n-1} \ge \psi_{n-1}^2 + 3$, $\alpha_n \ge \psi_n^2 + 2$, and $\beta_i = \tilde\beta_i$, $|\tilde\beta_i| \le |\psi_i + \psi_{i+1}|$, $1 \le i \le n-1$. Now we use the fact that formula (8) with T given by (4) can be expressed in the form

$$
\begin{aligned}
2v^T T(0) v &= [v_1, v_2] \begin{bmatrix} 2\alpha_1 & 2\beta_1 \\ 2\beta_1 & \alpha_2 - 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \\
&+ [v_2, v_3] \begin{bmatrix} \alpha_2 + 1 & 2\beta_2 \\ 2\beta_2 & \alpha_3 \end{bmatrix} \begin{bmatrix} v_2 \\ v_3 \end{bmatrix} \\
&+ \sum_{i=3}^{n-3} [v_i, v_{i+1}] \begin{bmatrix} \alpha_i & 2\beta_i \\ 2\beta_i & \alpha_{i+1} \end{bmatrix} \begin{bmatrix} v_i \\ v_{i+1} \end{bmatrix} \\
&+ [v_{n-2}, v_{n-1}] \begin{bmatrix} \alpha_{n-2} & 2\beta_{n-2} \\ 2\beta_{n-2} & \alpha_{n-1} + 1 \end{bmatrix} \begin{bmatrix} v_{n-2} \\ v_{n-1} \end{bmatrix} \\
&+ [v_{n-1}, v_n] \begin{bmatrix} \alpha_{n-1} - 1 & 2\beta_{n-1} \\ 2\beta_{n-1} & 2\alpha_n \end{bmatrix} \begin{bmatrix} v_{n-1} \\ v_n \end{bmatrix} \\
&\ge [v_1, v_2] \begin{bmatrix} 2(\psi_1^2 + 2) & 2\tilde\beta_1 \\ 2\tilde\beta_1 & \psi_2^2 + 2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \\
&+ \sum_{i=2}^{n-2} [v_i, v_{i+1}] \begin{bmatrix} \psi_i^2 + 4 & 2\tilde\beta_i \\ 2\tilde\beta_i & \psi_{i+1}^2 + 4 \end{bmatrix} \begin{bmatrix} v_i \\ v_{i+1} \end{bmatrix} \\
&+ [v_{n-1}, v_n] \begin{bmatrix} \psi_{n-1}^2 + 2 & 2\tilde\beta_{n-1} \\ 2\tilde\beta_{n-1} & 2(\psi_n^2 + 2) \end{bmatrix} \begin{bmatrix} v_{n-1} \\ v_n \end{bmatrix}. \qquad (9)
\end{aligned}
$$

Since

$$
\begin{aligned}
2(\psi_i^2 + 2)(\psi_{i+1}^2 + 2) - 4\beta_i^2 &\ge 2(\psi_i^2 + 2)(\psi_{i+1}^2 + 2) - 4(\psi_i + \psi_{i+1})^2 = \psi_i^2 \psi_{i+1}^2 + 8 - 8\psi_i\psi_{i+1} \\
&= 2(\psi_i\psi_{i+1} - 2)^2 \ge 0, \qquad i \in \{1, n-1\}, \\
(\psi_i^2 + 4)(\psi_{i+1}^2 + 4) - 4\beta_i^2 &\ge (\psi_i^2 + 4)(\psi_{i+1}^2 + 4) - 4(\psi_i + \psi_{i+1})^2 = \psi_i^2 \psi_{i+1}^2 + 16 - 8\psi_i\psi_{i+1} \\
&= (\psi_i\psi_{i+1} - 4)^2 \ge 0, \qquad 2 \le i \le n-2,
\end{aligned}
$$

all matrices used in the right hand side of (9) have positive diagonal elements and are positive semidefinite. If at least one of expressions $\psi_1\psi_2 - 2$, $\psi_i\psi_{i+1} - 4$, $2 \le i \le n-1$, $\psi_{n-1}\psi_n - 2$ is nonzero, the matrix $T(0)$ is positive definite by (a). Since eigenvalues of symmetric matrix depend continuously on its elements, the matrix $T(\varepsilon)$ is also positive definite, if number $\varepsilon > 0$ is sufficiently small. □

Theorem 4 can be used if the objective function is derived from a boundary value problem for ordinary differential equations.

Example 2 Consider a boundary value problem for the second order ordinary differential equation

$$y''(t) = \varphi(y(t)), \quad 0 \le t \le 1, \quad y(0) = y_0, \quad y(1) = y_1, \tag{10}$$

where function $\varphi : R \to R$ is twice continuously differentiable. If we divide the interval $[0, 1]$ onto $n + 1$ parts using nodes $t_i = ih$, $0 \le i \le n + 1$, where $h = 1/(n + 1)$ is the mesh-size and if we replace the second order derivatives in the nodes with differences

$$y''(t_i) = \frac{y(t_{i-1}) - 2y(t_i) + y(t_{i+1})}{h^2}, \quad 1 \le i \le n,$$

we obtain a system of n nonlinear equations

$$f_i(x) \triangleq h^2\varphi(x_i) + 2x_i - x_{i-1} - x_{i+1} = 0, \tag{11}$$

where $x_i = y(t_i)$, $0 \le 1 \le n + 1$, so $x_0 = y_0$ and $x_{n+1} = y_1$. Solving this system by the least squares method, the minimized function has the form

$$F(x) = \frac{1}{2}f^T(x)f(x) = \frac{1}{2}\sum_{i=1}^{n} f_i^2(x) = \frac{1}{2}\sum_{i=1}^{n}\left(h^2\varphi(x_i) + 2x_i - x_{i-1} - x_{i+1}\right)^2, \tag{12}$$

where $x = [x_1, \ldots, x_n]^T$ and $f = [f_1, \ldots, f_n]^T$. Differentiating only by x_{i-1}, x_i, x_{i+1}, we can write

$$\nabla f_i(x) = \begin{bmatrix} -1 \\ \psi(x_i) \\ -1 \end{bmatrix}, \qquad \nabla^2 f_i(x) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \psi'(x_i) & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

where $\psi(x_i) = 2+h^2\varphi'(x_i)$ and $\psi'(x_i) = h^2\varphi''(x_i)$. For a sum of squares, the Hessian matrix $G(x)$ can be expressed in the form $G(x) = J^T(x)J(x) + W(x)$, where $J(x)$ is the Jacobian matrix of mapping $f(x)$ and $W(x)$ is a second order term. To simplify the notation, we introduce these matrices for the particular case with $n = 5$. In this case, we can write

$$J(x) = \begin{bmatrix} \psi_1 & -1 & 0 & 0 & 0 \\ -1 & \psi_2 & -1 & 0 & 0 \\ 0 & -1 & \psi_3 & -1 & 0 \\ 0 & 0 & -1 & \psi_4 & -1 \\ 0 & 0 & 0 & -1 & \psi_5 \end{bmatrix}, \quad W(x) = \begin{bmatrix} f_1\psi_1' & 0 & 0 & 0 & 0 \\ 0 & f_2\psi_2' & 0 & 0 & 0 \\ 0 & 0 & f_3\psi_3' & 0 & 0 \\ 0 & 0 & 0 & f_4\psi_4' & 0 \\ 0 & 0 & 0 & 0 & f_5\psi_5' \end{bmatrix},$$

$$J^T(x)J(x) = \begin{bmatrix} \psi_1^2 + 1 & -(\psi_1 + \psi_2) & 1 & 0 & 0 \\ -(\psi_1 + \psi_2) & \psi_2^2 + 2 & -(\psi_2 + \psi_3) & 1 & 0 \\ 1 & -(\psi_2 + \psi_3) & \psi_3^2 + 2 & -(\psi_3 + \psi_4) & 1 \\ 0 & 1 & -(\psi_3 + \psi_4) & \psi_4^2 + 2 & -(\psi_4 + \psi_5) \\ 0 & 0 & 1 & -(\psi_4 + \psi_5) & \psi_5^2 + 1 \end{bmatrix},$$

where $\psi_i = \psi(x_i)$, $\psi_i' = \psi'(x_i)$, $1 \le i \le n$, which demonstrates that Hessian matrix $G(x) = J^T(x)J(x)+W(x)$ is pentadiagonal. If function $\varphi : R \to R$ is linear (so $\varphi'(x_i) = \varphi'$, $\varphi''(x_i) = 0$, $1 \le i \le n$, where φ' is the constant slope of linear function φ), one has

$W(x) = 0$, so $G(x) = J^T(x)J(x)$. Returning to the general case (with an arbitrary n) and assuming that function φ is linear, we can deduce that elements of the pentadiagonal matrix $G(x) = J^T(x)J(x)$ have the form (7) with inequalities replaced by equalities (this is the worst case). This matrix is pentadiagonal but not diagonally dominant. If $h^2|\varphi'| < 2$, we can write

$$\begin{aligned}\tilde{\alpha}_i - |\tilde{\gamma}_{i-2}| - |\tilde{\beta}_{i-1}| - |\tilde{\beta}_i| - |\tilde{\gamma}_i| &= \psi_i^2 + 2 - 2 - \psi_{i-1} - 2\psi_i - \psi_{i+1} \\ &= (2 + h^2\varphi')^2 - 4(2 + h^2\varphi') = (h^2\varphi')^2 - 4 < 0\end{aligned}$$

for $2 \le i \le n-1$ (where $\tilde{\gamma}_0 = \tilde{\gamma}_{n-1} = 0$). Therefore, we cannot use Theorem 3. At the same time, the linearity of function φ implies that $\psi_i = 2 + h^2\varphi'$, $1 \le i \le n$. If $(2 + h^2\varphi')^2 \ne 2$, one has $\psi_1\psi_2 - 2 \ne 0$, $\psi_{n-1}\psi_n - 2 \ne 0$, but if $(2 + h^2\varphi')^2 \ne 4$, one has $\psi_i\psi_{i+1} - 4$, $2 \le i \le n-1$. Thus assumptions of Theorem 4 are satisfied and matrix $T(0)$ (and also $T(\varepsilon)$, if number $\varepsilon > 0$ is sufficiently small) is positive definite.

Remark 2 If we are close to the solution, where $F(x) = 0$, then $f_i \approx 0$, $1 \le i \le n$, in (12). Moreover, absolute values of elements of the matrix $\mathrm{diag}(\psi_1', \ldots, \psi_n')$ are usually small in comparison with absolute values of elements of the matrix $J(x)^T J(x)$ (if $n \approx 1000$, then $h^2 \approx 10^{-6}$). Since a small change of diagonal elements does not violate the positive definiteness of matrix $T(0)$, we can expect that this matrix is positive definite in a sufficiently small neighborhood of the solution even if function $\varphi : R \to R$ in (10) is not linear. Then matrix $T(\varepsilon)$ corresponding to Example 2 is also positive definite if number $\varepsilon > 0$ is sufficiently small.

In Example 1, the Hessian matrix is an even order diagonally dominant Toeplitz matrix. It is interesting that for odd order diagonally dominant Toeplitz matrices this situation cannot appear. In the subsequent considerations, we denote elements of the Toeplitz matrix $G(x)$ by symbols c_i, $1 \le i \le n$. Thus

$$G(x) = \begin{bmatrix} c_1 & c_2 & c_3 & \ldots & c_{n-2} & c_{n-1} & c_n \\ c_2 & c_1 & c_2 & \ldots & c_{n-3} & c_{n-2} & c_{n-1} \\ c_3 & c_2 & c_1 & \ldots & c_{n-4} & c_{n-3} & c_{n-2} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ c_{n-2} & c_{n-3} & c_{n-4} & \ldots & c_1 & c_2 & c_3 \\ c_{n-3} & c_{n-2} & c_{n-3} & \ldots & c_2 & c_1 & c_2 \\ c_n & c_{n-1} & c_{n-2} & \ldots & c_3 & c_2 & c_1 \end{bmatrix}. \tag{13}$$

Theorem 5 *Let the Hessian matrix $G(x)$ be an odd order diagonally dominant Toeplitz matrix with positive diagonal elements. Then, if $\delta_i = \delta$, $1 \le i \le n$, and the number $\varepsilon > 0$ is sufficiently small, the matrix $T(\varepsilon)$ obtained by Remark 1 is positive definite.*

Proof (a) For n odd, equality $G(x)v_2 = T(0)v_2$ (see Remark 1) implies relations

$$\beta_1 = \sum_{j=1}^{\frac{n-1}{2}} c_{2j},$$

$$\begin{aligned}
\beta_1 + \beta_2 &= c_2 + \sum_{j=1}^{\frac{n-1}{2}} c_{2j} \quad \Rightarrow \quad \beta_2 = c_2, \\
\beta_2 + \beta_3 &= c_2 + \sum_{j=1}^{\frac{n-3}{2}} c_{2j} \quad \Rightarrow \quad \beta_3 = \sum_{j=1}^{\frac{n-3}{2}} c_{2j}, \\
\beta_3 + \beta_4 &= c_2 + c_4 + \sum_{j=1}^{\frac{n-3}{2}} c_{2j} \quad \Rightarrow \quad \beta_4 = c_2 + c_4, \\
\beta_4 + \beta_5 &= c_2 + c_4 + \sum_{j=1}^{\frac{n-5}{2}} c_{2j} \quad \Rightarrow \quad \beta_5 = \sum_{j=1}^{\frac{n-5}{2}} c_{2j},
\end{aligned}$$

etc., so

$$\beta_i = \sum_{j=1}^{\frac{n-i}{2}} c_{2j}, \quad \mathrm{mod}(j,2) = 1, \qquad \beta_i = \sum_{j=1}^{\frac{i}{2}} c_{2j}, \quad \mathrm{mod}(j,2) = 0.$$

For i odd, one has

$$|\beta_{i-1}| + |\beta_i| = \left| \sum_{j=1}^{\frac{i-1}{2}} c_{2j} \right| + \left| \sum_{j=1}^{\frac{n-i}{2}} c_{2j} \right| \leq \sum_{j=1}^{\frac{i-1}{2}} |c_{2j}| + \sum_{j=1}^{\frac{n-i}{2}} |c_{2j}|,$$

where the right-hand side contains the sum of absolute values of elements with even indices, which appear in the i-th row of matrix $G(x)$. The same result holds for i even.

(b) Equality $G(x)v_1 = T(0)v_1$ implies that the diagonal element α_i of matrix $T(0)$ is equal to the sum of elements with odd indices, which appear in the i-th row of matrix $G(x)$. Connecting this fact with the result introduced in (a), we can see that the number $\alpha_i - |\beta_{i-1}| - |\beta_i|$ is not greater than the number obtained by subtracting from c_1 the absolute values of all nondiagonal elements, appearing in the i-th row of the matrix $G(x)$. Since matrix $G(x)$ is diagonally dominant and $c_1 > 0$, this number is positive. □

In theorems concerning tridiagonal preconditioners obtained by numerical differentiation, we have assumed that all differences are the same, so $\delta_i = \delta$, $1 \leq i \leq n$. If $\tilde{\gamma}_i \geq 0$, $1 \leq i \leq n-2$, the assumptions of Theorem 3 can be substantially weakened. In this case, we can use arbitrary differences δ_i, $1 \leq i \leq n$ (e.g. $\delta_i = \varepsilon \max(|x_i|, 1)$, $1 \leq i \leq n$), and the Hessian matrix $G(x)$ may not be diagonally dominant).

Theorem 6 *Let the Hessian matrix $G(x)$ be pentadiagonal with nonnegative elements in the second off-diagonals (so $\tilde{\gamma}_i \geq 0$, $1 \leq i \leq n-2$). Let tridiagonal matrix, which arises from $G(x)$ after setting to zero these nonnegative elements, is diagonally dominant with positive diagonal elements. Ten, if δ_i, $1 \leq i \leq n$, are arbitrary and the number $\varepsilon > 0$ is sufficiently small, matrix $T(\varepsilon)$ obtained by Remark 1 is positive definite*

Proof If the differences δ_i, $1 \leq i \leq n$, are not the same, Theorem 2 implies that for $1 \leq i \leq n$ one has

$$\alpha_i = \tilde{\gamma}_{i-2} \frac{\delta_{i-2}}{\delta_i} + \tilde{\alpha}_i + \tilde{\gamma}_i \frac{\delta_{i-2}}{\delta_i}, \qquad \beta_{i-1} \frac{\delta_{i-1}}{\delta_i} + \frac{\delta_{i+1}}{\delta_i} \beta_i = \tilde{\beta}_{i-1} \frac{\delta_{i-1}}{\delta_i} + \frac{\delta_{i+1}}{\delta_i} \tilde{\beta}_i$$

(where $\tilde{\gamma}_{-1} = \tilde{\gamma}_0 = \tilde{\beta}_0 = \beta_0 = 0$, $\tilde{\gamma}_{n-1} = \tilde{\gamma}_n = \tilde{\beta}_n = \beta_n = 0$), so $\alpha_i \geq \tilde{\alpha}_i$, $1 \leq i \leq n$ and $\beta_i = \tilde{\beta}_i$, $1 \leq i \leq n-1$. From positive definiteness and diagonal dominance of tridiagonal matrix with elements $\tilde{\alpha}_i > 0$, $1 \leq i \leq n$, and $\tilde{\beta}_i$, $1 \leq i \leq n-1$, it follows that

$$\alpha_i - |\beta_{i-1}| - |\beta_i| \geq \tilde{\alpha}_i - |\tilde{\beta}_{i-1}| - |\tilde{\beta}_i| > 0$$

$1 \leq i \leq n$ (where $\tilde{\beta}_0 = \tilde{\beta}_n = 0$). □

In this section, we have demonstrated that the tridiagonal preconditioner determined by the numerical differentiation has a theoretical support and can be advantageously used for various types of problems. In the next section we show that this preconditioner is really efficient for solving practical problems

3 Implementation notes and numerical experiments

Since the truncated Newton PCG subalgorithm is stopped (in the fourth row) if the preconditioned Hessian matrix is not positive definite, we consider only positive definite preconditioners. Violation of positive definiteness can be detected during the Choleski or the Gill-Murray [11] decomposition procedure. Nevertheless, the tridiagonal matrix $T = T(\varepsilon)$ can be preliminary modified. We have tested the following possibilities:

- Matrix T is not modified (so $C = I$ if T is not positive definite).
- Diagonal elements of matrix T are replaced by their absolute values.
- Diagonal elements of matrix T are replaced by their absolute values and the off-diagonal elements are possibly changed (their absolute values are decreased) by the procedure described in [14].
- Matrix T is modified during the Gill-Murray decomposition to be positive definite.

The first two possibilities mentioned above gave approximately the same results, the third one was usually worse and the fourth possibility was quite unsuitable. Thus we will suppose, in the subsequent considerations and in the numerical experiments, that matrix T is not preliminary modified.

The basic tridiagonally preconditioned truncated Newton method, described in the previous section, considerably decreases the number of inner conjugate gradient iterations. Unfortunaleky the total number of gradient evaluations can be slightly greater in comparison with the unpreconditioned case (since two extra gradients are computed in every outer Newton iteration). This situation, which can be observed from performance profiles introduced below, arises if a small number of inner iterations suffices for both compared methods. To improve the performance profiles of the basic method, we can simply combine this method with the unpreconditioned truncated Newton method. The main idea is to use unpreconditioned iterations if their number is less than M and to switch to preconditioned iterations in the opposite case. Moreover, if the preconditioner $C = T$ is not positive definite, we turn to unpreconditioned iterations. This idea can be formally described in the following way.

(1) Set $L = 0$ and $M = 10$ in the first Newton iteration (the value $M = 10$ was obtained by numerical experiments).
(2) In all Newton iterations do:
   (a) If $L = 0$ set $C = I$ else compute tridiagonal matrix T by Remark 1 and set $C = T$.
   (b) If $L = 1$ and matrix C is not positive definite, set $C = I$ and $L = 0$
   (c) Determine the direction vector by the truncated Newton PCG subalgorithm.
   (d) If $L = 0$ and the number of conjugate gradient iterations in (c) was greater than M, set $L = 1$.

This method will be called the combined tridiagonally preconditioned truncated Newton method.

Now we are in the position to introduce the results of numerical experiments serving for the comparison of our tridiagonally preconditioned truncated Newton methods with the unpreconditioned truncated Newton method and the truncated Newton method that uses the limited memory preconditioner. The last method is based on the LBFGS updates proposed in [16]. We have used three LBFGS updates in every outer Newton iteration (it corresponds to the choice $m = 3$ in [16]). In fact we have tested also other preconditioners described in [14], but the results obtained did not bring new significant information. For testing truncated Newton methods, we have chosen three collections of large-scale unconstrained optimization problems. The first collection Test 11, described in [15], contains 58 test problems with 1000-5000 variables obtained from the CUTE collection [2] (we have used 54 problems solved by the unpreconditioned truncated Newton method). The second collection Test 12, described in [1], contains 73 test problems with 10000 variables (we have used 71 problems). The third collection Test 25, described in [13], contains 82 test problems with 1000 variables obtained from various sources (we have used 71 problems). Subroutines corresponding to the collections Test 11 and Test 25 can be found on `http://www.cs.cas.cz/luksan/test.html` (together with reports [13] and [15]) and subroutines corresponding to the collection Test 12 can be found on `http://camo.ici.ro/neculai/ansoft.htm`.

The summary results of computational experiments are reported in three tables corresponding to three collections Test 11, Test 12 and Test 25. The tables contain the following data: `NIT` – the total number of outer iterations, `NFV` – the total number of function evaluations, `NFG` – the total number of gradient evaluations, `NCGR` – the total number of inner iterations, `NIP` – the total number of preconditioned outer iterations, `TIME` – the total computational time. The rows correspond to the methods tested: `TN` – the unpreconditioned truncated Newton method, `TNTB` – the basic tridiagonally preconditioed truncated Newton method, `TNTC` – the combined tridiagonally preconditioed truncated Newton method, `TNLM` – the truncated Newton method preconditioned by the limited memory BFGS updates.

Method	NIT	NFV	NFG	NCGR	NIP	TIME
TN	6827	11071	364563	348768	-	33.55
TNTB	6805	11193	185827	156145	1621	18.95
TNTC	6742	10916	194394	175013	1031	20.18
TNLM	4945	10082	328568	315689	4945	39.34

Test 11 – 54 problems with 1000-5000 variables

Method	NIT	NFV	NFG	NCGR	NIP	TIME
TN	10674	13347	289817	270542	-	42.54
TNTB	8758	11500	78581	43622	2560	20.12
TNTC	9229	11895	62933	44503	138	12.63
TNLM	10824	12081	234372	217158	10824	69.02

Test 12 – 71 problems with 10000 variables

Method	NIT	NFV	NFG	NCGR	NIP	TIME
TN	7425	11826	372799	359516	-	23.25
TNTB	7631	12017	128887	99909	5661	9.54
TNTC	7220	11572	124049	99977	4994	9.47
TNLM	7262	12532	232408	219474	7262	15.97

Test 25 – 71 problems with 1000 variables

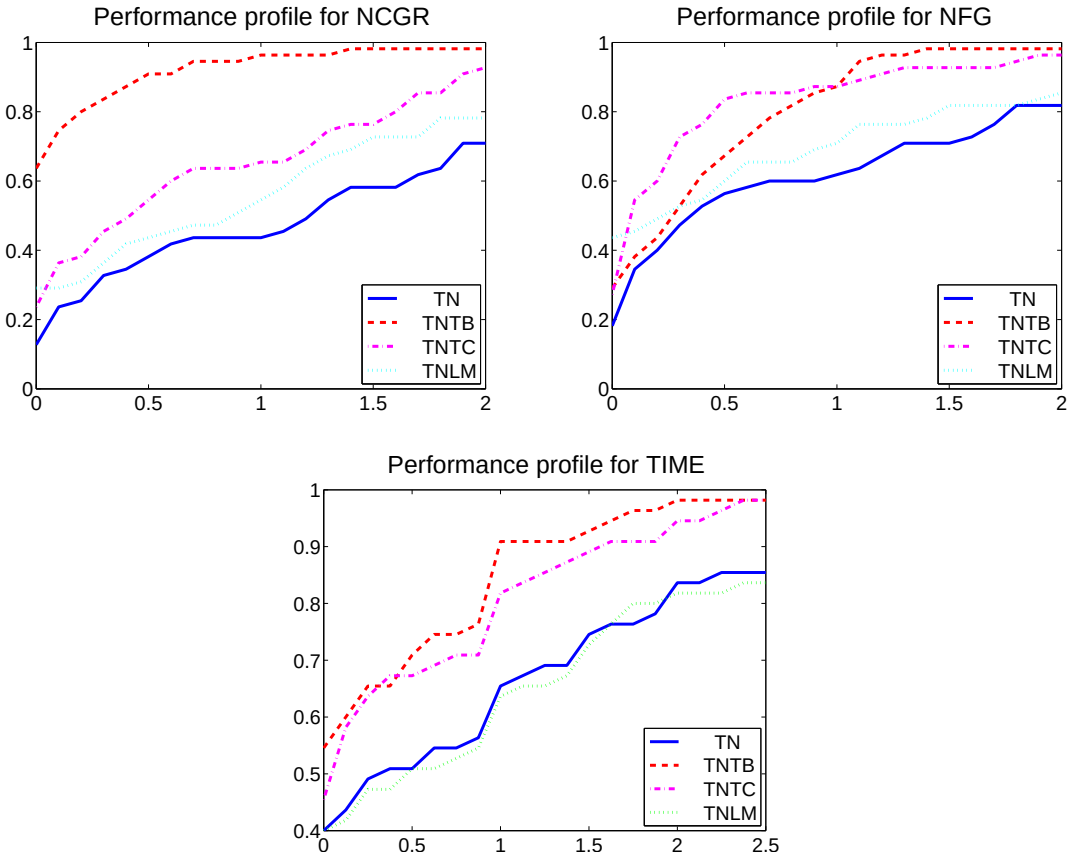
The results reported in the above tables imply several conclusions. First, the total number of gradient evaluations and also the total computational time are considerably less for methods TNTB and TNTC in comparison with methods TN and TNLM. Secondly, the number of the Newton iterations where preconditioner $C = T(\varepsilon)$ obtained by the method TNTB was used (since it was positive definite) is relatively large (about 1/4 Newton iterations for TEST11 and 3/4 Newton iterations for TEST25), so the use of such preconditioner is reasonable. Note, that we used all problems solved by the unpreconditioned method TN, so no selection of problems, which could be favourable for our preconditioners, was performed.

For a better demonstration of both the efficiency and the reliability, we compare the investigated truncated Newton methods by using performance profiles introduced in [6]. The performance profile $\pi_M(\tau)$ is defined by the formula

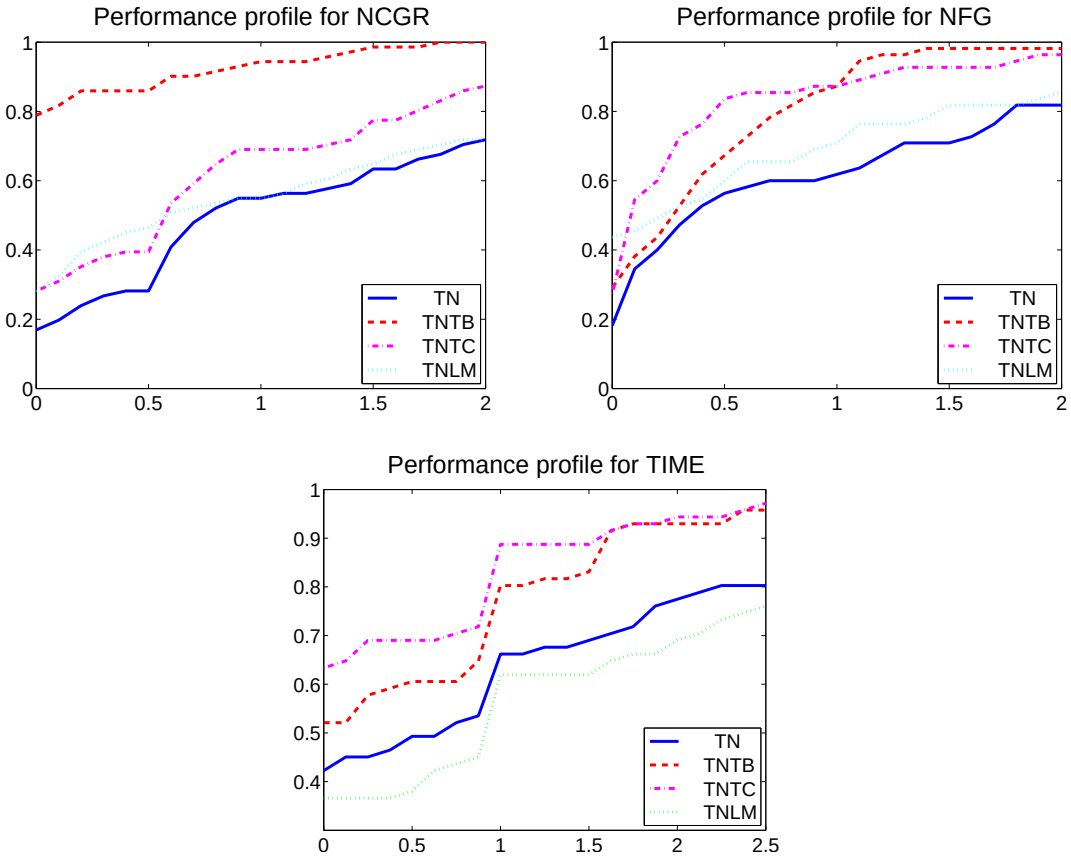
$$\pi_M(\tau) = \frac{\text{number of problems where } \log_2(\tau_{P,M}) \leq \tau}{\text{total number of problems}}$$

with $\tau \geq 0$, where $\tau_{P,M}$ is the performance ratio of the number of function evaluations (or the time) required to solve problem P by method M to the lowest number of function evaluations (or the time) required to solve problem P.

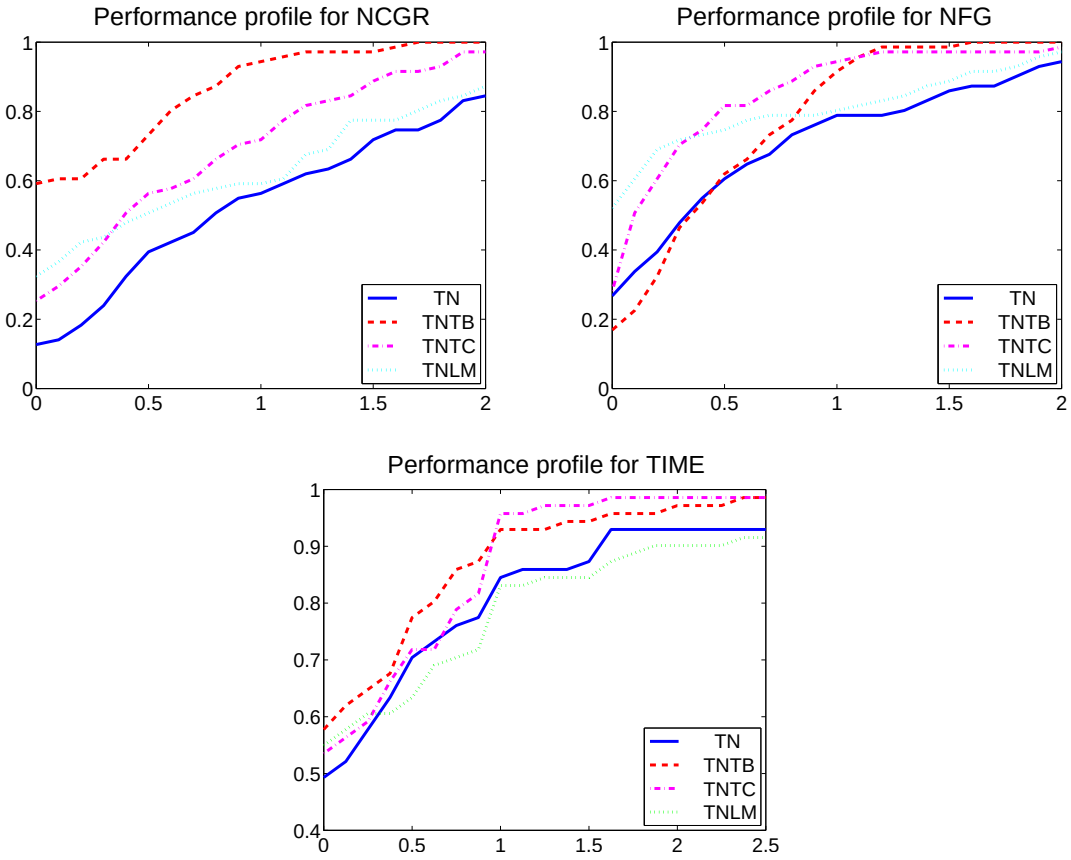
The value of $\pi_M(\tau)$ at $\tau = 0$ gives the percentage of test problems for which the method M is the best and the value for τ large enough is the percentage of test problems that method M can solve. The relative efficiency and reliability of each method can be directly seen from the performance profiles: the higher is the particular curve the better is the corresponding method. The following figures, reveal the performance profiles for tested methods graphically. In these figures `NCGR` and `NFG` are relative values (i.e. `NCGR`/`NIT` and `NFG`/`NIT`).



Test 11 – 54 problems with 1000-5000 variables



Test 12 – 71 problems with 10000 variables



Test 25 – 71 problems with 1000 variables

The above figures imply several conclusions. First, the efficiency of preconditioning, measured by the number of inner conjugate gradient iterations per one outer Newton iteration, is highest for the `TNTB` method. Secondly, since this methods computes two additional gradients in every Newton iteration, its efficiency, measured by the number of gradient evaluations per one Newton iteration, can be slightly worse in comparison with the unpreconditioned method. This deficiency can be eliminated using the `TNTC` method, which seems to be best, measured by the number of gradient evaluations per one Newton iteration.

References

[1] N. Andrei: An unconstrained optimization test functions collection, Advanced Modeling and Optimization **10** (2008) 147-161.

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

[3] A.R. Conn, N.I.M. Gould, P.L. Toint: *Trust-Region Methods* SIAM, Philadelphia, 2000.

[4] R.S. Dembo, S.C. Eisenstat, T. Steihaug: Inexact Newton methods. SIAM J. on Numerical Analysis **19** (1982) 400-408.

[5] R.S. Dembo, T. Steihaug: Truncated Newton algorithms for large-scale optimization. Math. Programming **26** (1983) 190-212.

[6] E.D. Dolan, J.J. Moré: Benchmarking optimization software with performance profiles. Mathematical Programming **91** (2002) 201-213.

[7] G Fasano: Planar-conjugate gradient algorithm for large scale unconstrained optimization. Part 1: Theory. Journal of Optimization Theory and Applications **125** (2005) 523-541.

[8] G Fasano: Planar-conjugate gradient algorithm for large scale unconstrained optimization. Part 2: Applications. Journal of Optimization Theory and Applications **125** (2005) 543-558.

[9] G. Fasano, M. Roma: Preconditioning Newton-Krylov methods in nonconvex large scale optimization. Report DIS 01-2007, Dipartimento di Informatica e Sistemistica "A. Ruberti" SAPIENZA - Universita di Roma, 2007.

[10] G. Fasano, M. Roma: AINVK: a class of approximate inverse preconditioners based on Krylov-subspace methods, for large indefinite linear systems.

[11] P.E. Gill, W. Murray: Newton type methods for unconstrained and linearly constrained optimization. Math. Programming, **7** (1974), 311-350.

[12] S. Gratton, A. Sartenaer, J. Tsimanga: On a class of limited memory preconditioners for large scale linear systems with multiple right-hand sides. SIAM J. on Optimization **21** (2011), 912-935.

[13] L. Lukšan, C. Matonoha, J. Vlček: Sparse test problems for unconstrained optimization. Report V-1064, Institute of Computer Science AS CR, Prague, 2010.

[14] L. Lukšan, C. Matonoha, J. Vlček: Band preconditioners for the matrix-free truncated Newton method. Report V-1079, Institute of Computer Science AS CR, Prague, 2010.

[15] L. Lukšan, C. Matonoha, J. Vlček: Sparse test problems for unconstrained optimization. Report V-1081, Institute of Computer Science AS CR, Prague, 2010.

[16] J.L. Morales, J. Nocedal: Automatic preconditioning by limited memory quasi-Newton updating. SIAM J. Optimization **10** (2000), 1079-1096.

[17] S.G. Nash: Newton-type minimization via Lanczos method. SIAM Journal on Numerical Analysis **21** (1984), 770-788.

[18] S.G. Nash: Preconditioning of truncated-Newton methods. SIAM Journal on Scientific and Statistical Computation **6** (1985), 599-616.

[19] J. Nocedal, S.J. Wright: *Numerical Optimization*. Springer, New York, 1999.

[20] D.P. O'Leary: A discrete Newton algorithm for minimizing a function of many variables. Mathematical Programming **23** (1983), 20-33.

[21] C.C. Paige, M.A. Saunders: Solution of sparse indefinite systems of linear equations. SIAM Journal on Numerical Analysis **12** (1975), 617-629.

[22] M. Roma: Dynamic scaling based preconditioning for truncated Newton methods in large scale unconstrained optimization. Optimization Methods and Software **20** (2005), 693-713.

[23] Y Saad: *Iterative Methods for Sparse Linear Systems*. SIAM, Philadelphia, 2003.

[24] T Steihaug: Damped inexact quasi-Newton methods. Report MASC TR 81-3, Department of Mathematical Sciences, Rice University, Houston, Texas 1984.

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

[26] 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.