Ladislav Lukšan; Ctirad Matonoha; Jan Vlček

Robust preconditioners for the matrix free truncated Newton method

In: Jan Chleboun and Petr Přikryl and Karel Segeth and Jakub Šístek (eds.): Programs and Algorithms of Numerical Mathematics, Proceedings of Seminar. Dolní Maxov, June 6-11, 2010. Institute of Mathematics AS CR, Prague, 2010. pp. 137--151.

Persistent URL: http://dml.cz/dmlcz/702752

ROBUST PRECONDITIONERS FOR THE MATRIX FREE TRUNCATED NEWTON METHOD*

Ladislav Lukšan, Ctirad Matonoha, Jan Vlček

Abstract

New positive definite preconditioners for the matrix free truncated Newton method are given. Corresponding algorithms are described in detail. Results of numerical experiments that confirm the efficiency and robustness of the preconditioned truncated Newton method are reported.

1 Introduction

We consider the unconstrained minimization problem

$$x^* = \arg\min_{x \in R^n} F(x), \quad F \in \mathcal{C}^2 : R^n \to R, \quad n - \text{large}$$

and use the notation

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

$$\|G(x)\| \leq \overline{G}, \quad \forall x \in R^n.$$

Numerical methods for unconstrained minimization are iterative and their iteration step has the form

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

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

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

for direction determination in such a way that

$$s_k = \arg\min_{s \in \mathcal{M}_k} Q(x_k + s).$$

There are two basic possibilities for direction determination: the line-search method, where

$$\mathcal{M}_k = R^n,$$

and the trust-region method, where

*This work was supported by the Czech Science Foundation, project No. 201/09/1957, and the institutional research plan No. AV0Z10300504.

$$\mathcal{M}_k = \{s \in R^n : \|s\| \le \Delta_k\}$$

(here $\Delta_k > 0$ is the trust region radius). We suppose that matrix $G = G(x)$ and its structure are not explicitly known. The direction vector (a minimum of a quadratic function) is in this case computed iteratively by the preconditioned conjugate gradient (PCG) method with preconditioner C. The outer index k is for the sake of simplicity mostly omitted.

Algorithm 1 *Direction determination by the PCG method (the line-search method).*
Data: Relative precision $0 \le \omega < 1$.

$s_1 = 0, \quad g_1 = g, \quad h_1 = C^{-1}g_1, \quad \rho_1 = g_1^T h_1, \quad p_1 = -h_1.$
Do $i = 1$ **to** m
$q_i = Gp_i, \quad \sigma_i = p_i^T q_i.$
If $\sigma_i \le 0$ **then** $s = s_i,$ stop.
$\alpha_i = \rho_i/\sigma_i, \quad s_{i+1} = s_i + \alpha_i p_i, \quad g_{i+1} = g_i + \alpha_i q_i,$
$h_{i+1} = C^{-1}g_{i+1}, \quad \rho_{i+1} = g_{i+1}^T h_{i+1}.$
If $\|g_{i+1}\| \le \omega\|g_1\|$ **or** $i = m$ **then** $s = s_i,$ stop.
$\beta_i = \rho_{i+1}/\rho_i, \quad p_{i+1} = -h_{i+1} + \beta_i p_i.$
End do

Algorithm 2 *Direction determination by the PCG method (the trust-region method)*
Data: Relative precision $0 \le \omega < 1$, *trust region radius* $\Delta > 0$.

$s_1 = 0, \quad g_1 = g, \quad h_1 = C^{-1}g_1, \quad \rho_1 = g_1^T h_1, \quad p_1 = -h_1.$
Do $i = 1$ **to** m
$q_i = Gp_i, \quad \sigma_i = p_i^T q_i.$
If $\sigma_i \le 0$ **then** $s = s_i + \lambda_i p_i,\ \lambda_i > 0,\ \|s_i + \lambda_i p_i\| = \Delta,$ stop.
$\alpha_i = \rho_i/\sigma_i.$
If $\|s_i + \alpha_i p_i\| \ge \Delta$ **then** $s = s_i + \lambda_i p_i,\ \lambda_i > 0,\ \|s_i + \lambda_i p_i\| = \Delta,$ stop.
$s_{i+1} = s_i + \alpha_i p_i, \quad g_{i+1} = g_i + \alpha_i q_i,$
$h_{i+1} = C^{-1}g_{i+1}, \quad \rho_{i+1} = g_{i+1}^T h_{i+1}.$
If $\|g_{i+1}\| \le \omega\|g_1\|$ **or** $i = m$ **then** $s = s_i,$ stop.
$\beta_i = \rho_{i+1}/\rho_i, \quad p_{i+1} = -h_{i+1} + \beta_i p_i.$
End do

Since matrix G is not given explicitly, we use numerical differentiation instead of matrix multiplication. Thus the product $q = Gp$ is replaced by the difference

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

where $\delta = \varepsilon/\|p\|$ (usually $\varepsilon = \sqrt{\varepsilon_M}$ and ε_M is a machine precision). The following theorems are proved in [4], Section 8.4.

Theorem 1 *Let function $F \in \mathcal{C}^2 : R^n \to R$ have Lipschitz continuous second order derivatives (with a 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

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

Theorem 2 *Consider the conjugate gradient method applied to the system of linear equations $G(x)s + g = 0$, where the vectors $q_i = G(x)p_i$ are replaced by the vectors $\tilde{q}_i = (g(x + \delta_i p_i) - g(x))/\delta_i$, $\delta_i = \varepsilon/\|p_i\|$. Suppose that the assumptions of Theorem 1 are satisfied and denote*

$$s_{m+1} = s_1 + \sum_{i=1}^{m} \alpha_i p_i, \quad g_{m+1} = g_1 + \sum_{i=1}^{m} \alpha_i q_i, \quad \tilde{g}_{m+1} = g_1 + \sum_{i=1}^{m} \alpha_i \tilde{q}_i$$

(thus $g_{m+1} = g + G(x)s_{m+1}$ if the computation is exact). Then it holds

$$\|\tilde{g}_{m+1} - g_{m+1}\| \leq \overline{\vartheta}\|s_{m+1}\|, \quad \overline{\vartheta} = \frac{m}{2}\varepsilon\overline{L}.$$

Remark 1 *Assume that $\|\tilde{g}_{m+1}\| \leq \overline{\omega}\|g\|$, $0 < \overline{\omega} < 1$, in the m-th step of the conjugate gradient method. If we set $s = s_{m+1}$ and $\tilde{g} = \tilde{g}_{m+1}$, then we can write*

$$\frac{\|\tilde{G}s + g\|}{\|g\|} \leq \overline{\omega}, \quad \frac{\|(\tilde{G} - G)s\|}{\|s\|} \leq \overline{\vartheta},$$

see Theorem 2, where $\tilde{G}$ is a symmetric matrix for which it holds $\tilde{G}s + g = \tilde{g}$ and $\overline{\vartheta} = m\varepsilon\overline{L}/2$. These expressions allow us to estimate the asymptotic rate of convergence.

A disadvantage of the difference version of the truncated Newton method 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. Therefore, the conjugate gradient method must be suitably preconditioned. Standard approaches cannot be used because matrix G is unknown. The following possibilities will be studied:

- Preconditioning based on the limited memory BFGS (Broyden, Fletcher, Goldfarb, Shanno) method.
- Band preconditioners obtained by the standard BFGS method equivalent to the preconditioned conjugate gradient method.
- Band preconditioners obtained by numerical differentiation.
- Tridiagonal preconditioners determined by the Lanczos method equivalent to the unpreconditioned conjugate gradient method.

2 Preconditioning based on the limited memory BFGS method

The idea of limited memory preconditioners is very simple (see [7]). Matrix $C_k^{-1} = H_k = H_k^k$, used as a preconditioner in the k-th step of the Newton method, is determined recurrently in such a way that $H_{k-l}^k = \gamma_{k-l} I$ where l is the number of updates (usually $l = 3$) and

$$\begin{aligned} H_{j+1}^k &= H_j^k + \left(\frac{y_j^T H_j^k y_j}{y_j^T d_j} + 1\right) \frac{d_j d_j^T}{y_j^T d_j} - \frac{H_j^k y_j d_j^T + d_j (H_j^k y_j)^T}{y_j^T d_j} \\ &= V_j^T H_j^k V_j + \frac{d_j d_j^T}{y_j^T d_j} \end{aligned}$$

for $k - l \le j \le k - 1$ with

$$V_j = I - \frac{y_j d_j^T}{y_j^T d_j}, \quad d_j = x_{j+1} - x_j, \quad y_j = g_{j+1} - g_j.$$

Matrix H_k is not computed explicitly. In the i-th inner step of the conjugate gradient method used in the k-th outer step of the Newton method, a vector $h_i = C_k^{-1} g_i = H_k g_i$ is determined by the Strang recurrences [6]. First, we set $u_k = g_i$ and compute numbers and vectors

$$\sigma_j = \frac{d_j^T u_{j+1}}{y_j^T d_j} \quad \text{and} \quad u_j = u_{j+1} - \sigma_j y_j, \quad k - l \le j \le k - 1,$$

respectively, using backward recurrences. Then we set $v_{k-l} = \gamma_{k-l} u_{k-l}$ and compute vectors

$$v_{j+1} = v_j + \left(\sigma_j - \frac{y_j^T v_j}{y_j^T d_j}\right) d_j, \quad k - l \le j \le k - 1,$$

using forward recurrence. Finally, we set $h_i = v_k$.

3 Band preconditioners obtained by the standard BFGS method

The BFGS method with perfect line search applied to a strictly convex quadratic function (with matrix G in the quadratic term) is equivalent to the conjugate gradient method with the same step-length choice. The BFGS method generates a sequence of matrices B_i, $1 \le i \le m$, in such a way that $B_1 = C$ and

$$B_{i+1} = B_i + \frac{y_i y_i^T}{d_i^T y_i} - \frac{B_i d_i (B_i d_i)^T}{d_i^T B_i d_i} = B_i + \frac{G p_i (G p_i)^T}{p_i^T G p_i} + \frac{g_i g_i^T}{p_i^T g_i}$$

for $1 \le i \le m$, where $d_i = s_{i+1} - s_i = \alpha_i p_i$ and $y_i = g_{i+1} - g_i = G d_i$. Vectors p_i and g_i are byproducts of the conjugate gradient method. If we use vectors $\tilde{q}_i$ (given by numerical differentiation) and $\tilde{g}_i$ instead of vectors $q_i = G p_i$ and g_i, respectively, we can write $B_1 = C$ and

$$B_{i+1} = B_i + \frac{\tilde{q}_i \tilde{q}_i^T}{p_i^T \tilde{q}_i} + \frac{\tilde{g}_i \tilde{g}_i^T}{p_i^T \tilde{g}_i}, \quad 1 \le i \le m.$$

From the above formulation, it is evident that only vectors generated by the preconditioned conjugate gradient method (with matrix multiplication replaced by numerical differentiation) are used for determination of matrices B_i, $1 \leq i \leq m$. These matrices do not occur in correction terms, so we can save only their selected parts (see [8]). If the vectors $\tilde{q}_i$ and $\tilde{g}_i$ are good approximations of the vectors q_i and g_i, then the matrices B_i, $1 \leq i \leq m$, are positive definite. Further, if the number of steps of the conjugate gradient method is sufficiently large, the matrix $B = B_{m+1}$ is a good approximation of matrix G so we can use it (or its part) as a preconditioner in the next step of the Newton method. We will investigate three special cases.

3.1 Diagonal preconditioning

If $C = D$, where D is a diagonal matrix containing diagonal elements of B, no problem arises because positive definite matrix B has positive numbers on the main diagonal. Diagonal preconditioning for problems with sparse Hessian matrices justifies the following theorem proved in [3].

Theorem 3 *Let $\mathcal{D}_n$ be the set of all diagonal matrices of order n and let D be a diagonal matrix containing diagonal elements of matrix G. Then it holds*

$$\kappa(GD^{-1}) \leq l \min_{M \in \mathcal{D}_n} \kappa(GM^{-1})$$

where κ is a spectral condition number and l is a maximal number of nonzero elements in rows of matrix G ($l = 5$ for pentadiagonal matrix G).

3.2 Tridiagonal preconditioning

Let now $C = T$ where T is a tridiagonal matrix containing elements of three main diagonals of matrix B. In this case the matrix C need not be positive definite (even if B was positive definite). Consider, as an example, matrices

$$B = \begin{bmatrix} 2 & -2 & 2 \\ -2 & 3 & -3 \\ 2 & -3 & 4 \end{bmatrix}, \qquad T = \begin{bmatrix} 2 & -2 & 0 \\ -2 & 3 & -3 \\ 0 & -3 & 4 \end{bmatrix}.$$

Both these matrices have positive elements on the main diagonal and positive main subdeterminants of the second order. But it holds that $\det B = 2$ and $\det T = -10$ so T is not positive definite, although B is positive definite. In order to remove this drawback, we have to modify matrix T to be positive definite.

Lemma 1 *Consider a tridiagonal matrix*

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

(elements α_i have differrent meaning than step-sizes α_i used in previous sections) and denote Δ_i a main subdeterminant of the i-th order of matrix T containing rows and columns with indexes $1, 2, \ldots, i$). Then it holds $\Delta_1 = \alpha_1$ and

$$\Delta_i = \alpha_i \Delta_{i-1} - \beta_{i-1}^2 \Delta_{i-2}, \qquad 2 \leq i \leq n,$$

where $\Delta_0 = 1$.

This well-known lemma can be used in the proof of the next theorem (see [1], [4]).

Theorem 4 *A tridiagonal matrix T is positive definite if and only if $\gamma_i > 0$ for $1 \leq i \leq n$, where $\gamma_1 = \alpha_1$ and*

$$\gamma_i = \alpha_i - \frac{\beta_{i-1}^2}{\gamma_{i-1}}, \qquad 2 \leq i \leq n.$$

Theorem 4 can be utilized in such a way that we compute numbers γ_i, $1 < i \leq n$, and as soon as $\gamma_i \leq 0$ for some index i, we decrease the off-diagonal element β_{i-1} so that $\beta_{i-1}^2 < \gamma_{i-1}\alpha_i$ (e.g. we set $\beta_{i-1}^2 = \lambda_{i-1}\gamma_{i-1}\alpha_i$, where $0 < \lambda_{i-1} < 1$). The trouble is that if we choose λ_{i-1} unsuitably, the resulting tridiagonal matrix can be ill-conditioned. For practical purposes it is more convenient to use the following theorem and its corollary (see [4]), Section 8.4.

Theorem 5 *Consider a tridiagonal matrix T with positive numbers on the main diagonal. If matrices*

$$\begin{bmatrix} 2\alpha_1 & 2\beta_1 \\ 2\beta_1 & \alpha_2 \end{bmatrix}, \qquad \begin{bmatrix} \alpha_i & 2\beta_i \\ 2\beta_i & \alpha_{i+1} \end{bmatrix}, \qquad \begin{bmatrix} \alpha_{n-1} & 2\beta_{n-1} \\ 2\beta_{n-1} & 2\alpha_n \end{bmatrix},$$

where $2 \leq i < n-2$, are positive semidefinite and at least one of them is positive definite, then matrix T is positive definite.

Corollary 1 *Let a tridiagonal matrix T contain the main diagonal and halves of subdiagonals of the positive definite matrix B (thus $\alpha_i = b_{i,i}$, $1 \leq i \leq n$, and $\beta_i = b_{i,i+1}/2$, $1 \leq i \leq n-1$). Then T is positive definite.*

Corollary 1 can be utilized so that the subdiagonal elements of matrix B are divided by two. Thereafter, the resulting tridiagonal matrix is positive definite. Theorem 5 can be utilized so that we compute determinants $\alpha_i\alpha_{i+1} - 4\beta_i^2$, $1 \leq i \leq n-1$, and as soon as $\alpha_i\alpha_{i+1} - 4\beta_i^2 < 0$ holds for some index i, we decrease the subdiagonal element β_i so that $\beta_i^2 = \alpha_i\alpha_{i+1}/4$.

3.3 Pentadiagonal preconditioning

Assertions of Theorem 5 and Corollary 1 can also be generalized for an arbitrary band matrix. We will show the corresponding procedure in case of the following pentadiagonal matrix

$$P = \begin{bmatrix} \alpha_1 & \beta_1 & \gamma_1 & \dots & 0 & 0 & 0 \\ \beta_1 & \alpha_2 & \beta_2 & \dots & 0 & 0 & 0 \\ \gamma_1 & \beta_2 & \alpha_3 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \alpha_{n-2} & \beta_{n-2} & \gamma_{n-2} \\ 0 & 0 & 0 & \dots & \beta_{n-2} & \alpha_{n-1} & \beta_{n-1} \\ 0 & 0 & 0 & \dots & \gamma_{n-2} & \beta_{n-1} & \alpha_n \end{bmatrix}$$

on which we will often refer. The following theorem and its corollary are proved in [4], Section 8.4.

Theorem 6 *Consider a pentadiagonal matrix P with positive elements on the main diagonal. If matrices*

$$\begin{bmatrix} \alpha_i & (3/2)\beta_i & 3\gamma_i \\ (3/2)\beta_i & \alpha_{i+1} & (3/2)\beta_{i+1} \\ 3\gamma_i & (3/2)\beta_{i+1} & \alpha_{i+2} \end{bmatrix}, \qquad 1 \le i < n-2,$$

are positive semidefinite, then matrix P is positive definite.

Corollary 2 *Let a pentadiagonal matrix P contain the main diagonal, two thirds of subdiagonals, and one third of subsubdiagonals of a positive definite matrix B (thus $\alpha_i = b_{i,i}$, $1 \le i \le n$, $\beta_i = 2b_{i,i+1}/3$, $1 \le i \le n-1$, and $\gamma_i = b_{i,i+2}/3$, $1 \le i \le n-2$). Then P is positive definite.*

Corollary 2 can be utilized so that we take two thirds of subdiagonal elements and one third of subsubdiagonal elements of matrix B. Thereafter, the resulting pentadiagonal matrix is positive definite. Theorem 6 can be utilized so that we first compute subdeterminants $\alpha_i\alpha_{i+1} - (9/4)\beta_i^2$, $1 \le i \le n-1$, and as soon as one of them is negative, we decrease the subdiagonal element β_i so that $\beta_i^2 = (4/9)\alpha_i\alpha_{i+1}$. Finally, we compute the determinants of the matrices mentioned in Theorem 6 as long as they are nonnegative. If one of them is negative, the corresponding element γ_i is modified using the following theorem is proved in [4], Section 8.4.

Theorem 7 *Determinants Δ_i of the matrices mentioned in Theorem 6 can be computed according to the formula*

$$\Delta_i = \alpha_{i+1}\left(\alpha_i\alpha_{i+2} - 9\gamma_i^2\right) - \frac{9}{4}\left(\alpha_i\beta_{i+1}^2 + \alpha_{i+2}\beta_i^2 - 6\beta_i\beta_{i+1}\gamma_i\right).$$

The determinant Δ_i is nonnegative if and only if $\underline{\gamma}_i \le \gamma_i \le \overline{\gamma}_i$ where

$$\begin{aligned}\underline{\gamma}_i &= \frac{1}{3\alpha_{i+1}}\left(\frac{9}{4}\beta_i\beta_{i+1} - \sqrt{D_i}\right),\\ \overline{\gamma}_i &= \frac{1}{3\alpha_{i+1}}\left(\frac{9}{4}\beta_i\beta_{i+1} + \sqrt{D_i}\right)\end{aligned}$$

are the roots of the quadratic equation $\Delta_i = 0$ *and*

$$D_i = \left(\alpha_i\alpha_{i+1} - \frac{9}{4}\beta_i^2\right)\left(\alpha_{i+1}\alpha_{i+2} - \frac{9}{4}\beta_{i+1}^2\right)$$

is the discriminant, divided by 36, of this equation, which is nonnegative provided that both multipliers are nonnegative.

Remark 2 *Theorem 7 offers two possibilities how to choose a new element* γ_i *in case that* $\Delta_i < 0$. *If* $\gamma_i < \underline{\gamma}_i$, *we set* $\gamma_i := \underline{\gamma}_i$. *If* $\gamma_i > \overline{\gamma}_i$, *we set* $\gamma_i := \overline{\gamma}_i$. *However, more advantageous is to set*

$$\gamma_i = \frac{1}{2}(\underline{\gamma}_i + \overline{\gamma}_i) = \frac{3}{4}\frac{\beta_i\beta_{i+1}}{\alpha_{i+1}},$$

because this choice is computationally simpler and gives better practical results.

4 Band preconditioners obtained by numerical differentiation

Suppose that the Hessian matrix has a band structure (even if it was not true in fact). The elements of this fictitious matrix that will be used as a preconditioner can be determined by numerical differentiation. It is performed only once at the beginning of the outer step of the Newton method.

In order to determine all elements of a band matrix which has $k-1$ couples of subdiagonals (thus $k = (l+1)/2$ where l is a band width), it suffices to use k gradient differences, which means to compute k extra gradients during each outer step of the Newton method. We will investigate three special cases again.

4.1 Diagonal preconditioning

Remark 3 *Assume that the Hessian matrix is diagonal. Then all its elements can be approximated using one gradient difference*

$$G(x)v \approx g(x+v) - g(x), \qquad v = [\delta_1, \ldots, \delta_n]^T,$$

where $\delta_1, \ldots, \delta_n$ *are suitable differences. Diagonal matrix* $C = D = \mathrm{diag}(\alpha_1, \ldots, \alpha_n)$ *where* $Dv = g(x+v) - g(x)$ *is then used as a preconditioner. After substitution we obtain* $\alpha_i\delta_i = g_i(x+v) - g_i(x)$ *or*

$$\alpha_i = \frac{g_i(x+v) - g_i(x)}{\delta_i}, \qquad 1 \le i \le n.$$

Remark 4 *The differences can be chosen in two different ways:*

(1) We set $\delta_i = \delta$, $1 \le i \le n$, *so* $v = \delta e$, *where* e *is a vector with all elements equal to one. We can choose (similarly as in Theorem 1)* $\delta = \sqrt{\varepsilon_M}/\|e\| = \sqrt{\varepsilon_M/n}$.

(2) We set $\delta_i = \sqrt{\varepsilon_M}\max(|x_i|, 1)$, $1 \le i \le n$. *This choice is less sensitive to rounding errors.*

In both cases we can write $\delta_i = \varepsilon\overline{\delta}_i$, $1 \le i \le n$, *where* $\varepsilon = \sqrt{\varepsilon_M}$ *and either* $\overline{\delta}_i = 1/\sqrt{n}$ *or* $\overline{\delta}_i = \max(|x_i|, 1)$ *for* $1 \le i \le n$.

A disadvantage of preconditioners based on numerical differentiation is the fact that they need not be positive definite. Consider a strictly convex quadratic function $F : R^2 \to R$:

$$F(x) = \frac{1}{2}\, x^T \begin{bmatrix} 1 & -2 \\ -2 & 6 \end{bmatrix} x, \qquad g(x) = \begin{bmatrix} 1 & -2 \\ -2 & 6 \end{bmatrix} x.$$

Then it holds

$$\frac{g(x+\delta e) - g(x)}{\delta} = \begin{bmatrix} 1 & -2 \\ -2 & 6 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 4 \end{bmatrix},$$

thus

$$De = \begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 4 \end{bmatrix},$$

which gives $\alpha_1 = -1$, $\alpha_2 = 4$, and so matrix D is not positive definite. This drawback can be removed by setting

$$\alpha_i = \frac{|g_i(x+v) - g_i(x)|}{\delta_i}, \qquad 1 \le i \le n.$$

This modification is justified by the following theorem proved in [10].

Theorem 8 *Let* $\mathcal{D}_n$ *be the set of all diagonal matrices of order* n *and let* $D = \mathrm{diag}(\alpha_1, \ldots, \alpha_n)$ *be a diagonal matrix such that*

$$\alpha_i = \sum_{j=1}^{n} |G_{ij}|, \qquad 1 \le j \le n,$$

where G_{ij}, $1 \le j \le n$, *are the elements of the* i*-th row of matrix* G. *Then it holds*

$$\kappa_1(GD^{-1}) = \min_{M \in \mathcal{D}_n} \kappa_1(GM^{-1}),$$

where κ_1 *is an* l_1 *condition number (the product of* l_1 *norms of a matrix and its inverse).*

If matrix G has only positive numbers and if we set $v = \delta e$, we can write $De = (g(x + \delta e) - g(x))/\delta \approx Ge$, so

$$\alpha_i \approx \sum_{j=1}^{n} G_{ij} = \sum_{j=1}^{n} |G_{ij}|$$

and matrix D is according to Theorem 8 an ideal preconditioner (in l_1 norm) for the system of equations $Gs + g = 0$. If matrix G does not contain only positive numbers, it holds

$$|\alpha_i| \approx \left| \sum_{j=1}^{n} G_{ij} \right| \leq \sum_{j=1}^{n} |G_{ij}|,$$

so the elements of modified matrix D form the lower bound for the elements of an ideal preconditioner.

4.2 Tridiagonal preconditioning

Theorem 9 *Let the Hesian matrix of function F be tridiagonal (as matrix T). 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 $\delta_i = \varepsilon\overline{\delta}_i$, $1 \leq i \leq n$. Then for $1 < i < n$ it holds*

$$\begin{aligned}
\alpha_1 &= \lim_{\varepsilon\to 0} \frac{g_1(x + v_1) - g_1(x)}{\delta_1}, \quad \beta_1 = \lim_{\varepsilon\to 0} \frac{g_1(x + v_2) - g_1(x)}{\delta_2}, \\
\alpha_i &= \lim_{\varepsilon\to 0} \frac{g_i(x + v_1) - g_i(x)}{\delta_i}, \quad \beta_i = \lim_{\varepsilon\to 0} \frac{g_i(x + v_2) - g_i(x) - \delta_{i-1}\beta_{i-1}}{\delta_{i+1}}, \quad \mathrm{mod}(i, 2) = 1, \\
\alpha_i &= \lim_{\varepsilon\to 0} \frac{g_i(x + v_2) - g_i(x)}{\delta_i}, \quad \beta_i = \lim_{\varepsilon\to 0} \frac{g_i(x + v_1) - g_i(x) - \delta_{i-1}\beta_{i-1}}{\delta_{i+1}}, \quad \mathrm{mod}(i, 2) = 0, \\
\alpha_n &= \lim_{\varepsilon\to 0} \frac{g_n(x + v_1) - g_n(x)}{\delta_n}, \quad \mathrm{mod}(n, 2) = 1, \\
\alpha_n &= \lim_{\varepsilon\to 0} \frac{g_n(x + v_2) - g_n(x)}{\delta_n}, \quad \mathrm{mod}(n, 2) = 0.
\end{aligned}$$

Remark 5 *Theorem 9, proved in [4], Section 8.4, gives us the way how to construct a tridiagonal preconditioner. A fixed number ε is chosen (e.g. $\varepsilon = \sqrt{\varepsilon_M}$) and the elements of matrix $C = T$ are computed according to formulas mentioned in Theorem 9 (in which the limit is omitted).*

Matrix $C = T$ obtained by Remark 5 need not be positive definite even if the Hessian matrix was positive definite. Tridiagonal matrix T obtained by application of Theorem 9 (with $\overline{\delta}_i = \overline{\delta}$, $1 \leq i \leq n$) to a strictly convex quadratic function of three variables with the positive definite Hessian matrix

$$G = \begin{bmatrix} 1 & -1 & -2 \\ -1 & 4 & -1 \\ -2 & -1 & 8 \end{bmatrix}$$

can serve as an example. We will state two theorems supporting a choice of tridiagonal preconditioning in cases when the actual Hessian matrix is pentadiagonal (see [4]).

Theorem 10 *Let the Hessian matrix $G(x)$ be pentadiagonal, positive definite, and diagonally dominant. Then, if $\delta_i = \varepsilon\overline{\delta}$, $1 \le i \le n$, and if the number ε is sufficiently small, matrix $C = T$ obtained by Remark 5 is positive definite and diagonally dominant.*

Remark 6 *Theorem 10, proved in [4], Section 8.4, requires all differences to be equal, which is fulfilled for instance when $\delta_i = \sqrt{2\varepsilon_M/n}$, $1 \le i \le n$. But the numerical experiments show that the choice $\delta_i = \sqrt{\varepsilon}\max(|x_i|, 1)$, $1 \le i \le n$, is usually more advantageous.*

Matrix T is positive definite for a lot of practical problems. Consider a boundary value problem for the second order ordinary differential equation

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

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 step-size and if we replace the second order derivatives in 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 will obtain a system of n nonlinear equations

$$h^2\varphi(x_i) + 2x_i - x_{i-1} - x_{i+1} = 0,$$

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

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

where $x = [x_1, \ldots, x_n]^T$. The following theorem is proved in [4], Section 8.4.

Theorem 11 *Let the difference version of the Newton method be applied to the sum of squares given above with a linear function $\varphi : R \to R$. Then, if $\delta_i = \varepsilon\overline{\delta}$, $1 \le i \le n$, and if the number ε is sufficiently small, matrix $C = T$ obtained by Remark 5 is positive definite.*

4.3 Pentadiagonal preconditioning

Theorem 12 *Let the Hessian matrix of function F be pentadiagonal (as matrix P). Set $v_1 = [\delta_1, 0, 0, \delta_4, 0, 0, \ldots]$, $v_2 = [0, \delta_2, 0, 0, \delta_5, 0, \ldots]$, $v_3 = [0, 0, \delta_3, 0, 0, \delta_6, \ldots]$, where $\delta_i = \varepsilon\overline{\delta}_i$, $1 \le i \le n$. Then it holds*

$$\alpha_i = \lim_{\varepsilon\to 0} \frac{g_i(x+v_1) - g_i(x)}{\delta_i}, \qquad \beta_i = \lim_{\varepsilon\to 0} \frac{g_i(x+v_2) - g_i(x) - \delta_{i-2}\gamma_{i-2}}{\delta_{i+1}},$$
$$\gamma_i = \lim_{\varepsilon\to 0} \frac{g_i(x+v_3) - g_i(x) - \delta_{i-1}\beta_{i-1}}{\delta_{i+2}}, \qquad \mathrm{mod}(i,3) = 1,$$
$$\alpha_i = \lim_{\varepsilon\to 0} \frac{g_i(x+v_2) - g_i(x)}{\delta_i}, \qquad \beta_i = \lim_{\varepsilon\to 0} \frac{g_i(x+v_3) - g_i(x) - \delta_{i-2}\gamma_{i-2}}{\delta_{i+1}},$$
$$\gamma_i = \lim_{\varepsilon\to 0} \frac{g_i(x+v_1) - g_i(x) - \delta_{i-1}\beta_{i-1}}{\delta_{i+2}}, \qquad \mathrm{mod}(i,3) = 2,$$
$$\alpha_i = \lim_{\varepsilon\to 0} \frac{g_i(x+v_3) - g_i(x)}{\delta_i}, \qquad \beta_i = \lim_{\varepsilon\to 0} \frac{g_i(x+v_1) - g_i(x) - \delta_{i-2}\gamma_{i-2}}{\delta_{i+1}},$$
$$\gamma_i = \lim_{\varepsilon\to 0} \frac{g_i(x+v_2) - g_i(x) - \delta_{i-1}\beta_{i-1}}{\delta_{i+2}}, \qquad \mathrm{mod}(i,3) = 0,$$

This theorem is proved in [4], Section 8.4.

5 Tridiagonal preconditioners determined by the Lanczos method

The elements of a tridiagonal matrix T obtained by the Lanczos method can be determined from the coefficients of the conjugate gradient method (which will be denoted with a tilde) by transformations $\alpha_1 = 1/\tilde{\alpha}_1$ and

$$\beta_i^2 = \frac{\tilde{\beta}_i}{\tilde{\alpha}_i^2}, \quad \alpha_{i+1} = \frac{\tilde{\beta}_i}{\tilde{\alpha}_i} + \frac{1}{\tilde{\alpha}_{i+1}}, \quad 1 \le i \le m,$$

where m is the number such that $\tilde{\alpha}_i > 0$ for $1 \le i \le m$. The following theorems are proved in [4], Section 8.4.

Theorem 13 *Consider the conjugate gradient method (applied to the quadratic function with the Hessian matrix G) such that $\tilde{\alpha}_i > 0$ for $1 \le i \le m$. Then the tridiagonal matrix T_m of order m with the elements given by the above transformations is positive definite.*

Remark 7 *The tridiagonal matrix T_m has the dimension $m \le n$. In order to obtain a preconditioner with the dimension n, we set*

$$C = [Q_m, Q_{n-m}] \begin{bmatrix} T_m & 0 \\ 0 & I_{n-m} \end{bmatrix} [Q_m, Q_{n-m}]^T = (I - Q_m Q_m^T) + Q_m T_m Q_m^T$$

where Q_m is a matrix with m orthonormal columns obtained with the symmetric Lanczos process and Q_{n-m} is a matrix with $n-m$ orthonormal columns such that matrix $[Q_m, Q_{n-m}]$ is square and orthogonal.

Theorem 14 *Let the assumptions of Theorem 13 be fulfilled. Then the preconditioner mentioned in Remark 7 is positive definite and it holds*

$$C^{-1} = (I - Q_m Q_m^T) + Q_m T_m^{-1} Q_m^T.$$

6 Rejecting of preconditioners

It is important to be able to decide whether the preconditioner will be used or rejected. Indefinite preconditioner is inappropriate also in case the Hessian matrix is not positive definite.

The Gill-Murray decomposition, proposed in [2], is a suitable means for testing positive definiteness and ill-conditioning of a matrix. If a pivot is during the elimination step less than $\delta \max(1, \max_{1 \leq i \leq n}(|\alpha_i|))$, where δ is a prescribed bound, then the decomposition of a preconditioner is terminated and the preconditioner is rejected. It is not worth performing the whole Gill-Murray decomposition and using the obtained positive definite matrix as a preconditioner (numerical experiments prove this claim). The number δ is usually chosen such that $\delta = 10^{-12}$. Sometimes, however, we have to choose a larger value (e.g. $\delta = 10^{-2}$).

7 Concluding remarks

- Preconditioning based on the limited memory BFGS method does not require any corrections. It is rather robust, but not very efficient.
- Band preconditioners obtained by the standard BFGS method have to be modified in advance, otherwise they are mostly rejected during the decomposition. Modifications based on Theorem 5, when the subdiagonal elements are decreased in order negative subdeterminants were zero, have proved to be very successful. It is shown that it is necessary to reject the preconditioners obtained in this way more often (e.g. to choose $\delta = 10^{-2}$).
- Band preconditioners obtained by numerical differentiation can be modified in a simple way that the diagonal elements are replaced with their absolute values. It suffices to choose $\delta = 10^{-12}$ for rejecting (except for diagonal preconditioners which are more prone to rejecting).
- It is not necessary to modify tridiagonal preconditioners determined by the Lanczos method (they are positive definite by Theorem 14). However, they can be determined only in unpreconditioned steps of the Newton method. This causes a lot of technical difficulties (the iteration process of the conjugate gradient method have to be modified).

8 Numerical comparison

The difference versions of the Newton method which use various preconditioners were tested using a set of 71 test problems with 1000 variables. The results are reported in the table containing the following data: `NIT` – the total number of iterations, `NFV` – the total number of function evaluations, `NFG` – the total number of gradient evaluations, `NCG` – the total number of inner iterations, `NCN` – the total number of preconditioned outer iterations, `NCP` – the total number of problems with enlarged bound for rejecting, `Time` – the total computational time.

The methods tested: `TN` – the unpreconditioned Newton method, `TNLM` – preconditioning using the limited memory BFGS method, `TNVM` – band preconditioning using the standard BFGS method (1 – diagonal, 2 – tridiagonal, 3 – pentadiagonal), `TNND` – band preconditioning using numerical differentiation (1 – diagonal, 2 – tridiagonal, 3 – pentadiagonal), `TNLT` – tridiagonal preconditioning using the Lanczos method, `LMVM` – the limited memory BFGS method, `CG` – the nonlinear conjugate gradient method. Methods `LMVM` and `CG` are mentioned only for comparison (they have nothing in common with the Newton method studied in this contribution).

Method	NIT	NFV	NFG	NCG	NCN	NCP	Time
`TN`	7425	11827	372789	359505	-	-	66.08
`TNLM`	7270	12521	233269	219347	7270	-	42.55
`TNVM-1`	7095	10303	274344	262855	4335	37	50.43
`TNVM-2`	6751	9252	139989	129933	4260	37	27.47
`TNVM-3`	6803	8857	229501	219820	4027	36	51.67
`TNND-1`	6522	8491	347384	331709	3857	40	59.51
`TNND-2`	7573	11245	147391	119434	4409	3	25.45
`TNND-3`	7107	10726	125262	91665	4943	4	24.57
`TNLT`	7398	11672	352199	339081	6808	1	55.61
`LMVM`	121314	127189	127189	-	-	-	39.59
`CG`	109166	325994	325994	-	-	-	75.72

From the results reported in this table, we can deduce several conclusions:

- The difference versions of the Newton method converge very fast, but they require more gradient computations.
- The unpreconditioned Newton method is not competitive with the limited memory BFGS method.
- Diagonal preconditioners and preconditioners obtained by the Lanczos method are not too efficient.
- Band preconditioners obtained by the standard BFGS method have to be often modified. Moreover, the bound for rejecting has to be often increased.
- Band preconditioners given by numerical differentiation rarely require corrections. The Newton method modified in this way is more efficient than the limited memory BFGS method.

References

[1] El-Mikkawy, M.E.A.: Notes on linear systems with positive definite tridiagonal matrices. Indian Journal on Pure and Applied Mathematics (2002), 1285–1293.

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

[3] Higham, N.J.: *Accuracy and stability of numerical algorithms*. SIAM, Philadelphia, 2002.

[4] Lukšan, L.: Numerické optimalizační metody. Tech. Rep. V-1058, Institute of Computer Science AS CR, Prague, 2009 (`www.cs.cas.cz/luksan/lekce4.pdf`).

[5] Lukšan, L., Matonoha, C., and Vlček, J.: Sparse test problems for unconstrained optimization. Tech. Rep. V-1064, Institute of Computer Science AS CR, Prague, 2010 (`ftp.cs.cas.cz/pub/reports/v1064-10.ps`).

[6] Matthies, H. and Strang, G.: The solution of nonlinear finite element equations. Int. J. for Numerical Methods in Engineering **14** (1979), 1613–1623.

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

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

[9] Nocedal, J.: Updating quasi-Newton matrices with limited storage. Mathematics of Computation **35** (1980), 773–782.

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