New Variable Metric Methods for Unconstrained Minimization Covering the Large-Scale Case

Vlček, Jan
2002

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

New variable metric methods for unconstrained minimization covering the large-scale case

J. Vlček, L. Lukšan

Technical report No. V 876

October 2002

Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic
phone: (+420)266052083 fax: (+420)286585789
e-mail: luksan@cs.cas.cz, vlcek@cs.cas.cz

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

New variable metric methods for unconstrained minimization covering the large-scale case

J. Vlček, L. Lukšan[1]

Technical report No. V 876
October 2002

Abstract

Some modifications and improvements of reduced-Hessian methods and a new family of numerically efficient variable metric or quasi-Newton methods for unconstrained minimization are given. These new methods give simple possibility of adaptation for large-scale optimization. Global convergence of the methods can be established for convex sufficiently smooth functions. Some encouraging numerical experience is reported.

Keywords

Unconstrained minimization, variable metric methods, limited-memory methods, global convergence, numerical results

[1]This work was supported by the grant No. 201/00/0080 given by the Czech Republic Grant Agency and with the subvention from Ministry of Education of the Czech Republic, project code MSM 242200002, L. Lukšan is also from Technical University of Liberec, Hálkova 6, 461 17 Liberec

1 Introduction

Variable metric (VM) methods, see [3], [9], for unconstrained minimization, are the most popular iterative methods for medium-size problems. Starting with an initial point $x_1 \in \mathcal{R}^N$, they generate a sequence $x_k \in \mathcal{R}^N$, $k \geq 1$, by the process $x_{k+1} = x_k + t_k d_k$, where $d_k \in \mathcal{R}^N$ is a direction vector and $t_k \geq 0$ is a stepsize.

Our original intention was to develop a limited-memory VM method for nonsmooth unconstrained optimization. We have tested many low storage methods, designed for the smooth case, see [5], [7], [13], [1], [11], but the results were disappointing. We were hardly able to solve any of the tested problems.

To test these methods better, we abandoned the nonsmooth case. From now on we assume that the problem function $f : \mathcal{R}^N \to \mathcal{R}$ has continuous second derivatives on the level set $\{x \in \mathcal{R}^N : \; f(x) \leq f(x_1)\}$ and denote $f_k = f(x_k)$, $g_k = \nabla f(x_k)$, $s_k = x_{k+1} - x_k$, $y_k = g_{k+1} - g_k$ and $\mathcal{G}_k = \mathrm{span}\{g_1, \dots, g_k\}$, $k \geq 1$.

In this paper we investigate the line search methods with

$$d_k = -H_k g_k, \qquad s_k = t_k d_k, \tag{1.1}$$

$k \geq 1$, where H_k is a symmetric positive definite matrix and the stepsize t_k is chosen in such a way that $t_k > 0$ and

$$f_{k+1} - f_k \leq \varepsilon_1 t_k g_k^T d_k, \qquad g_{k+1}^T d_k \geq \varepsilon_2 g_k^T d_k, \tag{1.2}$$

$k \geq 1$, where $0 < \varepsilon_1 < 1/2$ and $\varepsilon_1 < \varepsilon_2 < 1$.

The first important property of the line search method is the global convergence defined by relation

$$\liminf_{k \to \infty} |g_k| = 0. \tag{1.3}$$

The following theorem, see [3], [9], characterizes the global convergence of the line search method.

Theorem 1.1. *Let the objective function $f : \mathcal{R}^N \to \mathcal{R}$ be bounded from below and have bounded second derivatives. Consider the line search method satisfying (1.1)-(1.2). If*

$$\sum_{k=1}^{\infty} cos^2\theta_k \triangleq \sum_{k=1}^{\infty} \frac{(g_k^T H_k g_k)^2}{g_k^T g_k \, g_k^T H_k^2 g_k} = \infty, \tag{1.4}$$

then (1.3) holds.

The second important property of the line search method is the superlinear rate of convergence defined by relation

$$\lim_{k \to \infty} |x_{k+1} - x^\star| / |x_k - x^\star| = 0, \tag{1.5}$$

where $x^\star$ is the limit of the sequence $\{x_k\}_{k=1}^{\infty}$. The following theorem, see [2], [3], characterizes the superlinear rate of convergence of the line search method.

Theorem 1.2. *Consider the line search method satisfying (1.1)-(1.2) and such that $t_k = 1$ whenever this value fulfils (1.2). Let $x_k \to x^\star$, where $x^\star$ satisfies the second order sufficient conditions for the local minimum of f. If*

$$\lim_{k\to\infty} |(B_k - G_k)s_k|/|s_k| = 0, \tag{1.6}$$

where $G_k = G(x_k)$ is the Hessian matrix and $B_k = H_k^{-1}$, then an index $k_0 \ge 1$ exists such that $t_k = 1$, $k \ge k_0$, and $x_k \to x^\star$ superlinearly.

Condition (1.6) can also be written in another form. Since we can write $y_k = g_{k+1} - g_k = [\int_0^1 G(x_k + \xi s_k)d\xi]s_k$, $k \ge 1$, and since $x_k \to x^\star$ implies $s_k \to 0$, one has $|y_k - G_k s_k|/|s_k| \le \| \int_0^1 G(x_k + \xi s_k)d\xi - G(x_k)\| \to 0$ ($\|.\|$ denotes the spectral norm, unless explicitly indicated otherwise) and (1.6) is equivalent to

$$\lim_{k\to\infty} |B_k s_k - y_k|/|s_k| = 0. \tag{1.7}$$

We paid special attention to reduced-Hessian methods (e.g. [4], [5], [7], [17]) initially, because of some theoretical properties, significant for global convergence proof. We give some modifications and improvements in Section 2, but only briefly, because they affected our numerical results only insubstantially. During the seeking for a suitable limited-memory method we discovered a new family of VM methods, which we describe in Section 3. We call it the shifted Broyden family, because of its close relation to the well-known Broyden class, see e.g. [3]. We give the derivation of the new family, description of particular methods, the global convergence theory, some conditions for the superlinear rate of convergence and numerical results.

Section 4 is devoted to the related limited-memory methods. It contains theory, practical aspects, description of particular methods, the global convergence theory and numerical results.

2 Modifications of the reduced-Hessian method

2.1 Theoretical background

Let $\zeta > 0$, let A denote an $N \times N$ symmetric nonsingular matrix, let Z denote an $N \times m$ matrix, $m \le N$, such that matrix (Z, W) is orthogonal (which yields $Z^T Z = I$) for some $N \times (N - m)$ matrix W, and let $\mathcal{P}_Z = \text{range}(Z)$, $\mathcal{P}_Z^\perp = \text{null}(Z^T)$ and $\mathcal{A}_Z^\zeta = \{A :\ p \in \mathcal{P}_Z, q \in \mathcal{P}_Z^\perp \Rightarrow Ap \in \mathcal{P}_Z, Aq = \zeta q\}$. Each vector $p \in \mathcal{R}^N$ can be uniquely written as $p = p_Z + p_W$, where $p_Z \in \mathcal{P}_Z$, $p_W \in \mathcal{P}_Z^\perp$.

Lemma 2.1. *Let $p \in \mathcal{R}^N$, $q \in \mathcal{P}_Z$. Then $p_Z = ZZ^T p$, $p_W = (I - ZZ^T)p$, $ZZ^T q = q$.*

Proof. Let $q \in \mathcal{P}_Z$. Then $q = Zu$ for some $u \in \mathcal{R}^m$, thus $Z^T q = u$ and $q = ZZ^T q$. Let $p \in \mathcal{R}^N$. From $p = p_Z + p_W$ and since $p_Z \in \mathcal{P}_Z$ and $p_W \in \mathcal{P}_Z^\perp$, we have $ZZ^T p = ZZ^T p_Z = p_Z$, $p_W = p - p_Z = (I - ZZ^T)p$. □

Lemma 2.2. *The following properties of A are equivalent:*

$$(a)\ A \in \mathcal{A}_Z^\zeta, \qquad (b)\ A^{-1} \in \mathcal{A}_Z^{\zeta^{-1}}, \qquad (c)\ A = A_Z^\zeta,$$

where

$$A_Z^\zeta = ZZ^T AZZ^T + \zeta(I - ZZ^T). \tag{2.1}$$

If $A \in \mathcal{A}_Z^\zeta$ then the reduced matrix $Z^T AZ$ satisfies $(Z^T AZ)^{-1} = Z^T A^{-1} Z$.

Proof. $(a) \Rightarrow (c)$: Let $A \in \mathcal{A}_Z^\zeta$, $p \in \mathcal{R}^N$, $p = p_Z + p_W$. Then $A_Z^\zeta p = ZZ^T Ap_Z + \zeta p_W = Ap_Z + Ap_W = Ap$ by (2.1) and Lemma 2.1, thus $A_Z^\zeta = A$.

$(c) \Rightarrow (b)$: Let $A = A_Z^\zeta$. Using (2.1), we obtain $AZ = ZZ^T AZ$, or $Z = A^{-1}ZZ^T AZ$, thus $I = Z^T Z = (Z^T A^{-1} Z)(Z^T AZ)$. Therefore matrix $Z^T AZ$ is nonsingular and $(Z^T AZ)^{-1} = Z^T A^{-1} Z$ holds. Moreover, the relation $Z = A^{-1}ZZ^T AZ$ implies $A^{-1}Z = Z(Z^T AZ)^{-1}$ and if $p \in \mathcal{P}_Z$ then one has $p = Zu$ for some $u \in \mathcal{R}^m$, which yields $A^{-1}p = A^{-1}Zu = Z(Z^T AZ)^{-1}u \in \mathcal{P}_Z$. If $q \in \mathcal{P}_Z^\perp$, then $Aq = \zeta q$ by (2.1), thus $A^{-1}q = \zeta^{-1}q$.

$(b) \Rightarrow (a)$: Since we have proved $(a) \Rightarrow (b)$, it suffices to replace A by A^{-1}. $\square$

Theorem 2.1. *Let $\gamma\zeta > 0$, $\delta_i \in \mathcal{R}$, $p_i \in \mathcal{P}_Z$, $q_i \in \mathcal{P}_Z$, $i = 1, \ldots, n$, $n \geq 1$ and suppose that the matrix $P = \sum_{i=1}^n \delta_i p_i q_i^T$ is symmetric, $\gamma A + P$ is nonsingular and $A \in \mathcal{A}_Z^\zeta$. Then $\gamma A + P \in \mathcal{A}_Z^{\gamma\zeta}$.*

Proof. One has $(\gamma A+P)_Z^{\gamma\zeta} = ZZ^T(\gamma A+P)ZZ^T + \gamma\zeta(I-ZZ^T) = \gamma A_Z^\zeta + \sum_{i=1}^n \delta_i p_i q_i^T = \gamma A + P$ by (2.1) and Lemma 2.1, thus $\gamma A + P \in \mathcal{A}_Z^{\gamma\zeta}$ by Lemma 2.2. $\square$

Theorem 2.2. *Let Q be an orthogonal $m \times m$ matrix, let Z' be an $N \times m'$ matrix, $m' \leq N$, such that $\mathcal{P}_{Z'} \supset \mathcal{P}_Z$ and $(Z')^T Z' = I$ holds. Then $\mathcal{A}_{ZQ}^\zeta = \mathcal{A}_Z^\zeta \subset \mathcal{A}_{Z'}^\zeta$.*

Proof. The first relation follows from (2.1), Lemma 2.2 and $(ZQ)(ZQ)^T = ZZ^T$. Let $p \in \mathcal{P}_{Z'}$, $p = p_Z + p_W$, $q \in \mathcal{P}_{Z'}^\perp$ and $A \in \mathcal{A}_Z^\zeta$. Then $q \in \mathcal{P}_Z^\perp$ and $Aq = \zeta q$. Since $\mathcal{P}_Z \subset \mathcal{P}_{Z'}$, we have $Ap = Ap_Z + \zeta p_W \in \mathcal{P}_{Z'}$ by $A \in \mathcal{A}_Z^\zeta$, thus $A \in \mathcal{A}_{Z'}^\zeta$. $\square$

Theorem 2.3. *Let Q be an orthogonal $N \times N$ matrix and $A \in \mathcal{A}_Z^\zeta$. Then $QAQ^T \in \mathcal{A}_{QZ}^\zeta$.*

Proof. By (2.1) one has $(QAQ^T)_{QZ}^\zeta = QZZ^T Q^T QAQ^T QZZ^T Q^T + \zeta(I - QZZ^T Q^T) = QAQ^T$, thus $QAQ^T \in \mathcal{A}_{QZ}^\zeta$ by Lemma 2.2. $\square$

Utilizing this general theory, we denote by index k relevant quantities in iteration k. In the principal variant of the reduced-Hessian method, see e.g. [4], the subspaces $\mathcal{P}_{Z_k}$ and $\mathcal{G}_k$ are identical for every k.

Suppose that the initial VM matrix is $H_1 = \zeta_1 I \in \mathcal{A}_{Z_1}^{\zeta_1}$ and that $H_k \in \mathcal{A}_{Z_k}^{\zeta_k}$. In iteration k, we first replace Z_k by some Z_{k+1}, $\mathcal{P}_{Z_{k+1}} \supset \mathcal{P}_{Z_k}$, which yields $H_k \in \mathcal{A}_{Z_{k+1}}^{\zeta_k}$ by Theorem 2.2. Then we apply some Broyden VM update $H_k \to H_{k+1}$ (see [3], [9]), which has the form $H_{k+1} = \gamma_k(H_k + \sum_{i=1}^{r_k} \nu_i p_i p_i^T)$, $\gamma_k > 0$, $\nu_i \in \mathcal{R}$, $p_i \in \mathcal{P}_{Z_{k+1}}$, $1 \leq i \leq k$ (every update can contain together with any vector p also $H_k p$ or $H_k^{-1}p$ by definition $\mathcal{A}_{Z_{k+1}}^{\zeta_k}$ and Lemma 2.2). Setting $\zeta_{k+1} = \gamma_k \zeta_k$, one has $H_{k+1} \in \mathcal{A}_{Z_{k+1}}^{\zeta_{k+1}}$ by Theorem 2.1, thus always $H_k \in \mathcal{A}_{Z_k}^{\zeta_k}$, $k \geq 1$.

This property $H \in \mathcal{A}_Z^\zeta$ (omitting index k) is important, because then we can **equivalently** replace H by $H_Z^\zeta = Z(Z^T HZ)Z^T + \zeta(I - ZZ^T)$ in all computations by Lemma 2.2, thus we can proceed with the reduced matrix $Z^T HZ$ instead of H so that we have all iterates the same (in the precise arithmetic). Moreover, we see from equality $Z^T(H+\delta pq^T)Z = Z^T HZ + \delta(Z^T p)(Z^T q)^T$ that we can simply update reduced matrix $Z^T HZ$ using reduced vectors $Z^T p$, $Z^T q$ instead of updating matrix H using vectors p, q.

2.2 Matrix damage caused by discarding some basis vector

The situation will be quite different in the limited-memory version of the reduced-Hessian method. We suppose that $A \in \mathcal{A}_Z^\zeta$ and that we need to discard some column z of the basis matrix $Z = (\underline{Z}, z)$. Note that Z is usually multiplied from the right by some orthogonal matrix in advance (to adapt $\underline{Z}$ to stored vectors g_i, or better to s_i, see [5]), but this fact is not significant here and has no influence on validity $A \in \mathcal{A}_Z^\zeta$ by Theorem 2.2.

Usually, only the reduced matrix $\underline{Z}^T A \underline{Z}$ is formed, but we will investigate a modification of matrix $A \to \tilde{A}$ caused by the discarding of column z, with $\tilde{A} \in \mathcal{A}_{\underline{Z}}^\zeta$, to be able to utilize general theory. Naturally, we assume $\underline{Z}^T \tilde{A} \underline{Z} = \underline{Z}^T A \underline{Z}$. Then one has $\tilde{A} = \tilde{A}_{\underline{Z}}^\zeta = \underline{Z}\,\underline{Z}^T \tilde{A} \underline{Z}\,\underline{Z}^T + \zeta(I - \underline{Z}\,\underline{Z}^T) = A_{\underline{Z}}^\zeta$ by Lemma 2.2 . Note that for $A = \zeta I + \sum_{i=1}^r \delta_i q_i q_i^T$, $r \geq 1$, the replacement $A \to A_{\underline{Z}}^\zeta$ corresponds to the **projection** $q_i \to \underline{Z}\,\underline{Z}^T q_i$, $i \geq 1$. The following theorems describe properties of matrix $\tilde{A}$.

Theorem 2.4. *Let $\zeta > 0$ and suppose that A is positive definite. Then matrix $\tilde{A} = A_{\underline{Z}}^\zeta$ is positive definite and $\tilde{A} \in \mathcal{A}_{\underline{Z}}^\zeta$. If $p \in \mathcal{P}_{\underline{Z}}$, then $p^T \tilde{A} p = p^T A p$.*

Proof. It follows from equality $\underline{Z}^T \underline{Z} = I$ that $\tilde{A}_{\underline{Z}}^\zeta = (A_{\underline{Z}}^\zeta)_{\underline{Z}}^\zeta = A_{\underline{Z}}^\zeta = \tilde{A}$, which implies $\tilde{A} \in \mathcal{A}_{\underline{Z}}^\zeta$ by Lemma 2.2. Let $p \in \mathcal{R}^N$, $p = p_{\underline{Z}} + p_{\underline{W}}$, $p_{\underline{Z}} \in \mathcal{P}_{\underline{Z}}$, $p_{\underline{W}} \in \mathcal{P}_{\underline{Z}}^\perp$. We obtain $p^T \tilde{A} p = p^T \underline{Z}\,\underline{Z}^T A \underline{Z}\,\underline{Z}^T p + \zeta p^T (I - \underline{Z}\,\underline{Z}^T) p = p_{\underline{Z}}^T A p_{\underline{Z}} + \zeta(p_{\underline{Z}} + p_{\underline{W}})^T p_{\underline{W}} = p_{\underline{Z}}^T A p_{\underline{Z}} + \zeta p_{\underline{W}}^T p_{\underline{W}}$ by Lemma 2.1, which for $p \in \mathcal{P}_{\underline{Z}}$ (i.e. $p_{\underline{W}} = 0$) yields the desired equality. Let $p \neq 0$. Then the positive definiteness of $\tilde{A}$ follows from $p^T \tilde{A} p \geq p_{\underline{Z}}^T A p_{\underline{Z}} > 0$ for $p_{\underline{Z}} \neq 0$ and from $p^T \tilde{A} p = \zeta p_{\underline{W}}^T p_{\underline{W}} > 0$ for $p_{\underline{Z}} = 0$, i.e. $p_{\underline{W}} \neq 0$. □

Theorem 2.5. *Let $\zeta > 0$, $A \in \mathcal{A}_Z^\zeta$, $Z = (\underline{Z}, z)$ and $\tilde{A} = A_{\underline{Z}}^\zeta$. Then*

$$A - \tilde{A} = wz^T + zw^T + (\alpha - \zeta)zz^T, \tag{2.2}$$

where $w = \underline{Z}\,\underline{Z}^T Az$, $\alpha = z^T Az$. Moreover, one has $\mathrm{Tr}(A - \tilde{A}) = \alpha - \zeta$ and $\|A - \tilde{A}\|_F^2 = (\alpha - \zeta)^2 + 2|w|^2$ (Frobenius matrix norm).

Proof. By Lemma 2.2 one has $A = ZZ^T AZZ^T + \zeta(I - ZZ^T)$ and using $ZZ^T = \underline{Z}\,\underline{Z}^T + zz^T$ gives (2.2). The relation $\mathrm{Tr}(A - \tilde{A}) = \alpha - \zeta$ follows from $Z^T Z = I$, which implies $w^T z = 0$ and $|z| = 1$. Observe that we obtain further $\|A - \tilde{A}\|_F^2 = \|wz^T + zw^T\|_F^2 + (\alpha - \zeta)^2 |z|^4 = 2|w|^2 + (\alpha - \zeta)^2$. □

Note that we tested some possibilities of decreasing this matrix damage, without substantial improvement of the results.

2.3 Basis vector adding strategies

Usually, the new basis vector $\bar{z}$ is formed and added to Z in iteration k, if $|g_W| > \varepsilon_A |g|$, where we write $g = g_{k+1}$ and $\varepsilon_A > 0$ is an adding tolerance (typically $\varepsilon_A = 10^{-4}$). Then we set $\bar{z} = g_W / |g_W|$. The main disadvantage is that the new vector can be left out, while the old ones remain unchanged in the basis (sometimes even in many consecutive iterations).

One way out from this situation is represented by the following strategy: when $|g_W| \le \varepsilon_A |g|$, we sometimes discard some column z of $Z = (\underline{Z}, z)$ in advance to have value $|g_{\underline{W}}|$ sufficiently large, where $|g_{\underline{W}}| = |g - g_{\underline{Z}}|$. In view of $\underline{Z}\,\underline{Z}^T = ZZ^T - zz^T$ and $z \in \mathcal{P}_Z$ we easily obtain $|g_{\underline{W}}|^2 = |g - \underline{Z}\,\underline{Z}^T g|^2 = |g_W + (z^T g) z|^2 = |g_W|^2 + (z^T g)^2$, which can be advantageously utilized for the choice of z.

The following method of basis vector adding seems to be more hopeful, because we always add the new vector to the basis. First we set $\bar{z} = g_{k+1}/|g_{k+1}|$ and determine an orthogonal matrix Q as a product of plane rotations, for which vector $(ZQ)^T \bar{z}$ has the first $m-1$ elements equal to zero. Denoting by z'_i, $i = 1, \ldots, m$ the columns of ZQ, we then set $z''_m = z'_m - (\bar{z}^T z'_m)\bar{z} \perp \bar{z}$ and $\bar{Z} = (z'_1, \ldots, z'_{m-1}, z''_m / |z''_m|, \bar{z})$ for $z''_m \neq 0$, $\bar{Z} = (z'_1, \ldots, z'_{m-1}, \bar{z})$ otherwise. Obviously $(ZQ)^T ZQ = I = \bar{Z}^T \bar{Z}$ and $\mathcal{P}_{\bar{Z}} \supset \mathcal{P}_{ZQ} = \mathcal{P}_Z$ and we can pass from basis Z to $\bar{Z}$ by Theorem 2.2. Note that in practice we leave out z''_m not only when $z''_m = 0$, but also when $|z''_m| \le \varepsilon_A$, similarly as we leave out g_{k+1} in the usual method of basis vector adding.

Surprisingly, the plane rotations caused extreme growth of rounding errors here; these errors were approximately the same, when we replaced the plane rotations by Householder transformations, see [6].

2.4 Basis vector discarding using QR transformation

The choice of basis vectors to discard in this method should increase stability. The discarded vectors are replaced by their projection (see Section 2.2). Since the algorithm respecting the error analysis is rather complicated, we present only a simplified version. Note that there are other ways how to increase stability, but this is very robust.

Let $H = \zeta I + UMU^T$, $U = (u_1, \ldots, u_m)$, $m \ge 1$, rank $U > 0$. Initially, we have $M = \operatorname{diag}(\xi_1, \ldots, \nu_m)$, where $u_1, \ldots, u_m$ and $\xi_1, \ldots, \nu_m$ are computed using standard VM updates, see [3], [9]. Using QR transformation (e.g. Householder transformation with pivoting, see [6]), we can write

$$U = Q \begin{pmatrix} R & C_1 \\ 0 & C_2 \end{pmatrix}, \tag{2.3}$$

where Q is $N \times N$ orthogonal matrix, R is $r \times r$ nonsingular upper triangular matrix, whose diagonal elements are arranged in descending order (which minimizes column norms of C_2, i.e. discarding errors, see below) and $r \in [1, m]$ is chosen so that $(N-r) \times (m-r)$ matrix C_2 could be neglected. Denoting $\underline{U} = Q(R^T, 0)^T$ and $P = R^{-1} C_1$, one has $Q(C_1^T, 0)^T = Q(R^T, 0)^T R^{-1} C_1 = \underline{U} P$. Assuming $C_2 = 0$, we obtain $U = \underline{U}(I, P)$ by (2.3), thus we can reduce H to the form $H = \zeta I + \underline{U}\,\underline{M}\,\underline{U}^T$, where $\underline{M} = (I, P) M (I, P)^T$, and continue in VM updating, which does not change this form of matrix H representation, as we see from relation $H + \nu\, uu^T = \zeta I + (\underline{U}, u) \operatorname{diag}(\underline{M}, \nu)\, (\underline{U}, u)^T$.

We show that this neglecting C_2 corresponds to the vector projection, caused by some basis vector discarding, as was shown in Section 2.2. First we can define $\underline{Z} = \underline{U}R^{-1} = Q(I,0)^T$, since obviously range($\underline{U}$) = range($\underline{Z}$) and $\underline{Z}^T\underline{Z} = (I,0)Q^TQ(I,0)^T = I$. By (2.3), we now can write the corresponding projection in the form

$$\underline{Z}\,\underline{Z}^T U = Q \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} Q^T U = Q \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} R & C_1 \\ 0 & C_2 \end{pmatrix} = Q \begin{pmatrix} R & C_1 \\ 0 & 0 \end{pmatrix}.$$

The advantages of this method are easy computing of discarding errors and good stability, the disadvantage is the greater number of arithmetic operations in comparison with the reduced-Hessian method; this number can be reduced using a suitable strategy of choice r. Unfortunately, although the method can minimize the discarding errors, these errors were in practice very soon too great to be neglected, i.e. numerical results were not substantially better than in the reduced-Hessian method.

2.5 Methods without basis vector discarding

These methods are similar to the reduced-Hessian method except that the basis vector discarding is replaced by an orthogonal transformation, which preserves VM matrices eigenvalues and a certain number of direction vectors. We present only two versions: the first one preserves the maximum number of these vectors and appears to be more efficient, the second one preserves only the latest direction vector and seems to be more advantageous for small number of basis vectors.

Let $H \in \mathcal{A}_Z^\zeta$ (see Section 2.1), where $\zeta > 0$, $Z = (z_1, \dots, z_m)$, $m \le N$. Initially, we simply process the reduced-Hessian method until we need discard some basis vector. Let $g = g_{k+1}$ and $s = s_k$ be the latest values of gradient and basic points increment (in iteration k).

In the first version, we further suppose that we have the last increment vectors matrix (indices of vectors s_i are changed) $S = (s_1, \dots, s_m)$ such that $Z^T S$ is an upper triangular matrix; this property can be easily achieved, see e.g. [7]. First we replace matrix Z by $Z' = ZQ_1 = (z'_1, \dots, z'_m) \triangleq (z'_1, \underline{Z})$ such that $s_j^T z'_1 = 0$ (thus $s_j \in \mathcal{P}_{\underline{Z}}$) and $s_j^T z'_i = 0$, $1 < j < i \le m$, where Q_1 is an orthogonal matrix, product of plane rotations (the first row of $Z^T S$ is combined with the other ones). In this connection we correct the reduced matrix according to relation $(ZQ_1)^T H (ZQ_1) = Q_1^T (Z^T H Z) Q_1$.

For $g_{\underline{W}} \neq 0$, where $g_{\underline{W}} = g - \underline{Z}\,\underline{Z}g$, we then set $\bar{z} = g_{\underline{W}}/|g_{\underline{W}}|$, $\bar{z} = z'_1$ otherwise. Further we set $v = z'_1 - \bar{z}$ and $Q_2 = I - 2vv^T/|v|^2$ for $v \neq 0$, $Q_2 = I$ otherwise, and replace H by $\bar{H} = Q_2 H Q_2^T$; it can be achieved by replacing Z' by $Q_2 Z'$, without changing the reduced matrix $(Z')^T H Z' = (Q_2 Z')^T (Q_2 H Q_2^T)(Q_2 Z')$. Obviously, it holds $Q_2 \underline{Z} = \underline{Z}$ in view of $\underline{Z}^T v = 0$ (thus also $Q_2 s_j = s_j$, $j > 1$). Combining this with

$$Q_2 z'_1 = z'_1 - 2\frac{(z'_1 - \bar{z})^T z'_1}{|z'_1 - \bar{z}|^2}(z'_1 - \bar{z}) = z'_1 - 2\frac{1 - \bar{z}^T z'_1}{2 - 2\bar{z}^T z'_1}(z'_1 - \bar{z}) = \bar{z},$$

one has $Q_2 Z' = (\bar{z}, z'_2, \dots, z'_m)$. Lastly we replace $Q_2 Z'$ by $Z_+ = (z'_2, \dots, z'_m, \bar{z}) = Q_2 Z' Q_3$ for some orthogonal Q_3 and again correct the reduced matrix. It is easy to see that $\bar{H} \in \mathcal{A}_{Z_+}^\zeta$ by Theorem 2.2 and Theorem 2.3.

In the second version, we only set $\bar{H} = QHQ^T$ and $Z_+ = QZ$, where Q is an orthogonal matrix, such that $g \in \mathcal{P}_{Z_+}$, $Qs = s$ and that the angle between g and Qg is minimized, see Lemma 2.3 (if $g \in \mathcal{P}_Z$, the choice $Q = I$ is suitable). Again we have $\bar{H} \in \mathcal{A}^{\zeta}_{Z_+}$ by Theorem 2.3.

Finally, in both these versions, we update $\bar{H}$ to H_+, or equivalently $Z_+^T \bar{H} Z_+$ to $Z_+^T H_+ Z_+$ (see Section 2.1). Obviously, for update belonging to the Broyden class, see [3], [9], one has $H_+ \in \mathcal{A}^{\zeta}_{Z_+}$ by (1.1) and Theorem 2.1 and we can go to the next iteration.

Lemma 2.3. *Let $s \in \mathcal{P}_Z$, $s \neq 0$, $g \notin \mathcal{P}_Z$ and $Q = I - 2vv^T/|v|^2$, where $v = g_W - \theta w_Z$ with $w_Z = g_Z - (g^T s/|s|^2)s$ and $\theta = |g_W|^2/\left(|w_Z|^2 + |w_Z|\sqrt{|w_Z|^2 + |g_W|^2}\right)$ for $w_Z \neq 0$, $v = g - \alpha s - \beta \hat{s}$ otherwise, where $\beta^2 = (|s|^2|g|^2 - (s^T g)^2)/(|s|^2|\hat{s}|^2 - (s^T \hat{s})^2)$, $\alpha = (s^T g - \beta s^T \hat{s})/|s|^2$ and $\hat{s} \in \mathcal{P}_Z$ is some vector, linearly independent of s.*

Then orthogonal matrix Q satisfies $Qs = s$, $g \in \mathcal{P}_{QZ}$ and v maximizes quantity $g^T Qg/|g|^2$ subject to these conditions.

Proof. The condition $g \in \mathcal{P}_{QZ}$ is equivalent to $Qg = Q^T g \in \mathcal{P}_Z$. Let $Qg = q$ for some $q \in \mathcal{P}_Z$. Then $|q| = |Qg| = |g|$ and $q = g - 2(g^T v/|v|^2)v$. Since $q \neq g$ by $g \notin \mathcal{P}_Z$, it must be $g^T v \neq 0$ and v is proportional to $g - q$; we can set $v = g - q$. On the other hand, let $v = g - q$ for some $q \in \mathcal{P}_Z$ such that $|q| = |g|$. It is easy to see that $Qg = q$.

The condition $Qs = s$ is equivalent to $v^T s = 0$; a general solution of this equation can be written as $v = M(g - p)$, $M = I - ss^T/|s|^2$, $p \in \mathcal{R}^N$. Denoting $u = Mp$ and $w = Mg$, this yields $v = w - u$. Since $q = g - v = u + g - w$ and $w - g = Mg - g$ is proportional to s, one has $u \in \mathcal{P}_Z$. Observing that $Ms = 0$, this gives $0 = u^T s = w^T s$, thus $0 = u^T(w - g) = w^T(w - g)$. Combining it with $q = u - (w - g)$, we get $|q|^2 = |u|^2 + |w - g|^2 = |u|^2 + (|w - g|^2 + |w|^2) - |w|^2 = |u|^2 + |g|^2 - |w|^2$. Consequently, the condition $|q| = |g|$ is equivalent to $|u| = |w|$.

We want to minimize $2|g|^2(1 - g^T Qg/|g|^2) = 2g^T(g - q) = |g - q|^2 = |u - w|^2$ under the conditions $g \in \mathcal{P}_{QZ}$ and $Qs = s$ examined above. By $u \in \mathcal{P}_Z$ and $|u| = |w|$ we obtain $|u - w|^2 = |u|^2 - 2u^T w_Z + |w_Z|^2 + |w_W|^2 = (|w| - |w_Z|)^2 + 2(|u||w_Z| - u^T w_Z) + |w_W|^2$, which is for $w_Z \neq 0$ minimized, when u is proportional to w_Z, i.e. $u = (|w|/|w_Z|)w_Z$ by $|u| = |w|$. Since $w_Z = Mg_Z = g_Z - (g^T s/|s|^2)s$ and $w_W = g_W$, we obtain $v = w - u = w_W + w_Z - (|w|/|w_Z|)w_Z = g_W - \theta w_Z$ with $\theta = |w|/|w_Z| - 1 = (\sqrt{|w_Z|^2 + |g_W|^2} - |w_Z|)/|w_Z|$, which can be rewritten in the desired form.

If $w_Z = 0$, the quantity $|u - w| = \sqrt{2}|g_W|$ is independent of u or q. If we then set $v = g - q = g - \alpha s - \beta \hat{s}$, the conditions $|q| = |g|$ and $v^T s = 0$, equivalent to $g \in \mathcal{P}_{QZ}$ and $Qs = s$, give the desired relations (since we have two conditions, we need two parameters; note that we cannot choose g_Z as $\hat{s}$, because $w_Z = 0$ implies that g_Z is proportional to s). $\square$

It is interesting that numerical results were comparable with the reduced-Hessian method, in spite of the VM matrix damage caused by the orthogonal transformation.

3 Shifted variable metric methods

Variable metric methods, see [3], [9], use symmetric positive definite matrices H_k, $k \geq 1$; usually $H_1 = I$ and H_{k+1} is obtained from $\gamma_k H_k$ ($\gamma_k > 0$ is a scaling parameter) by a rank-two VM update to satisfy the quasi-Newton condition (in generalized form) $H_{k+1} y_k = \varrho_k s_k$, where $\varrho_k > 0$ is a nonquadratic correction parameter (see [9]).

In shifted VM methods, matrices H_k have the form

$$H_k = \zeta_k I + A_k, \tag{3.1}$$

$k \geq 1$, where $\zeta_k > 0$ and A_k are symmetric positive semidefinite matrices; usually $A_1 = 0$ and A_{k+1} is obtained from $\gamma_k A_k$ by a rank-two VM update to satisfy the shifted quasi-Newton condition

$$A_{k+1} y_k = \varrho_k \tilde{s}_k, \quad \zeta_{k+1} = \varrho_k \sigma_k, \tag{3.2}$$

where

$$\tilde{s}_k = s_k - \sigma_k y_k \tag{3.3}$$

and $\sigma_k > 0$ is a shift parameter. Obviously, relations (3.1)-(3.3) imply that matrix H_{k+1} satisfies the quasi-Newton condition $H_{k+1} y_k = \varrho_k s_k$.

In the subsequent analysis we use the following notation

$$a_k = y_k^T H_k y_k, \;\; \bar{a} = y_k^T A_k y_k, \;\; \hat{a}_k = y_k^T y_k, \;\; b_k = s_k^T y_k, \;\; \tilde{b}_k = \tilde{s}_k^T y_k, \;\; B_k = H_k^{-1},$$

$k \geq 1$. To simplify the notation we frequently omit index k and replace index $k+1$ by symbol $+$. Although we use the unit values of γ_k and ϱ_k in almost all cases, we will consider also non-unit values in the subsequent analysis as it is usual in case of VM methods (see [9]).

In this section we concentrate on shifted analogy of the Broyden class, see [3], [9], which we call the shifted Broyden family. Involving the scaling and the nonquadratic correction and using the same argumentation as in standard VM methods, we can write the shifted VM update for $\tilde{b} > 0$ (which implies $\tilde{s} \neq 0$, $y \neq 0$) in the form

$$\frac{1}{\gamma} A_+ = A + \frac{\varrho}{\gamma} \frac{\tilde{s}\tilde{s}^T}{\tilde{b}} - \frac{A y y^T A}{\bar{a}} + \frac{\eta}{\bar{a}} \left(\frac{\bar{a}}{\tilde{b}} \tilde{s} - A y \right) \left(\frac{\bar{a}}{\tilde{b}} \tilde{s} - A y \right)^T \tag{3.4}$$

(if $\bar{a} = 0$, i.e. $Ay = 0$, we simply omit the last two terms, because their limit value is zero for $Ay = \lim_{\xi \to 0} \xi q$, $\bar{a} = \lim_{\xi \to 0} \xi q^T y$, $q^T y \neq 0$), where η is a free parameter (verification of $A_+ y = \varrho \tilde{s}$ for this update is straightforward). There are two important special cases. For $\eta = 0$ we obtain the shifted DFP update, for $\eta = 1$ the shifted BFGS update

$$\frac{1}{\gamma} A_+^{DFP} = A + \frac{\varrho}{\gamma} \frac{\tilde{s}\tilde{s}^T}{\tilde{b}} - \frac{A y y^T A}{\bar{a}}, \quad \frac{1}{\gamma} A_+^{BFGS} = A + \left(\frac{\varrho}{\gamma} + \frac{\bar{a}}{\tilde{b}} \right) \frac{\tilde{s}\tilde{s}^T}{\tilde{b}} - \frac{\tilde{s} y^T A + A y \tilde{s}^T}{\tilde{b}}.$$

3.1 Basic properties

Theorem 3.1. *Let A be positive semidefinite, $\eta \geq 0$ and $\sigma\hat{a} < b$. Then matrix A_+ given by (3.4) is positive semidefinite.*

Proof. Since $\sigma\hat{a} < b$, relation (3.3) implies $\tilde{b} = \tilde{s}^T y = b - \sigma\hat{a} > 0$ and the positive semidefiniteness of matrix A_+ follows for $\bar{a} = 0$ from (3.4), otherwise from the quasi-product form of (3.4)

$$\frac{1}{\gamma}A_+ = \left(I - \left(\frac{\sqrt{\eta}}{\tilde{b}}\tilde{s} + \frac{1-\sqrt{\eta}}{\bar{a}}Ay\right)y^T\right)A\left(I - y\left(\frac{\sqrt{\eta}}{\tilde{b}}\tilde{s} + \frac{1-\sqrt{\eta}}{\bar{a}}Ay\right)^T\right) + \frac{\varrho}{\gamma}\frac{\tilde{s}\tilde{s}^T}{\tilde{b}}, \quad (3.5)$$

which can be readily verified, using straightforward arrangements and comparing corresponding terms. $\square$

Note that there are other useful quasi-product forms of (3.4), e.g.

$$\frac{1}{\gamma}A_+ = \left(I - py^T\right) A \left(I - yp^T\right) + \frac{\varrho}{\gamma}\frac{\eta\tilde{b}}{\omega^2\bar{a}^2}Ayy^TA, \quad (3.6)$$
$$p = \frac{\omega}{\tilde{b}}\tilde{s} + \frac{1}{\bar{a}}\left(1 - \frac{\eta}{\omega}\right)Ay, \qquad \omega = \pm\sqrt{\eta + (\varrho/\gamma)\tilde{b}/\bar{a}},$$

which becomes a product form for $\eta = 0$ and which can also be easily verified.

From now on we will suppose that $\eta \geq 0$. In view of Theorem 3.1, the shift parameter should satisfy inequality $0 < \sigma < b/\hat{a}$. Therefore, it is advantageous to introduce relative shift parameter $\mu = \sigma\hat{a}/b \in (0,1)$ and by (3.3) we can write

$$\sigma = \mu b/\hat{a}, \qquad \tilde{b} = \tilde{s}^T y = b - \sigma\hat{a} = b(1-\mu). \quad (3.7)$$

Note that if we set $\gamma = \zeta_+/\zeta$ (this case is however not so efficient as that with $\gamma = 1$) and use Woodbury formula $(H + UMU^T)^{-1} = B - BU(M^{-1} + U^TBU)^{-1}U^TB$ as in [9], we can derive update relation for B_+ from (3.4), because then by (3.1)

$$\frac{1}{\gamma}H_+ = \frac{1}{\gamma}(\zeta_+ I + A_+) = \zeta I + \frac{1}{\gamma}A_+ = H + \left(\frac{1}{\gamma}A_+ - A\right).$$

3.2 Determination of the shift parameter

Determination of the shift parameter σ (or μ) is a crucial part of the shifted VM method. Since $\zeta_+ = \varrho\sigma$ by (3.2), the choice of σ influences the lowest eigenvalue of matrix H_+. Therefore σ should not be close to zero when matrix A is not sufficiently positive definite. On the other hand, the norm of A_+ can increase explosively when σ tends to $b/\hat{a}$ (see below).

In the simplest shift parameter determination strategy the value of μ remains the same in all iterations. The values from the interval

$$0.20 \leq \mu \leq 0.25 \quad (3.8)$$

(e.g. the choice $\mu = 0.22$) are suitable in this case. If $\mu \geq 1/2$, then the convergence is usually lost (the shifted DFP method is an exception). In spite of the fact that we do not know all causes of this phenomenon, our following restricted analysis of the shifted BFGS method with $A = UU^T$, where U is a rectangular matrix, gives a useful formula for determination of parameter μ.

Lemma 3.1. *Denoting* $\nu = \mu/(1-\mu)$, $\phi = \nu\sqrt{1 - b^2/(\hat{a}|s|^2)}$, $V = I - sy^T/b$ *and* $\tilde{V} = I - \tilde{s}y^T/\tilde{b}$, *there holds* $\|\tilde{V} - V\|/\|V\| = \phi$. *Moreover, let vector* $u \in \mathcal{R}^N$, $y^T u \neq 0$, *be scaled to satisfy* $y^T u = b$. *Then*

$$\phi - \frac{|u-s|}{|u|}(1+\phi) \le \frac{|\tilde{V}u|}{|u|} \le \phi + \frac{|u-s|}{|u|}(1+\phi). \tag{3.9}$$

Proof. One has

$$\tilde{V} - V = \frac{sy^T - (\mu b/\hat{a})yy^T - (1-\mu)sy^T}{b(1-\mu)} = \nu\left(\frac{sy^T}{b} - \frac{yy^T}{\hat{a}}\right) = \frac{\nu}{b}\left(s - \frac{b}{\hat{a}}y\right)y^T$$

by (3.3) and (3.7). Observing that $b^2 \le \hat{a}|s|^2$ by the Schwartz inequality and that $\nu^2|s - (b/\hat{a})y|^2 = \nu^2(|s|^2 - b^2/\hat{a}) = |s|^2\phi^2$, this implies

$$\|\tilde{V} - V\|^2 = \|(\tilde{V} - V)^T(\tilde{V} - V)\| = (\nu/b)^2\,|s - (b/\hat{a})y|^2\,\|yy^T\| = \phi^2\hat{a}|s|^2/b^2.$$

Matrix $V^T V$ has one zero eigenvalue, $N-2$ unit eigenvalues and $\mathrm{Tr}(V^T V) = N - 2 + \hat{a}|s|^2/b^2$. Thus $\|V\|^2 = \hat{a}|s|^2/b^2$, which yields the first assertion.

Let $y^T u = b$. By (3.3) and (3.7) we get $\tilde{V}u = u - \tilde{s}/(1-\mu) = u - s - \nu[s - (b/\hat{a})y]$. Since we have $\nu|s - (b/\hat{a})y| = \phi|s|$, the rest follows from inequalities

$$\begin{aligned} |\tilde{V}u| &\le \phi|s| + |u-s| \le \phi(|u| + |u-s|) + |u-s| = \phi|u| + (1+\phi)|u-s|, \\ |\tilde{V}u| &\ge \phi|s| - |u-s| \ge \phi(|u| - |u-s|) - |u-s| = \phi|u| - (1+\phi)|u-s|. \qquad \square \end{aligned}$$

Now we turn back to the shift parameter determination. Value $\|\tilde{V} - V\|/\|V\|$, equal to ϕ by Lemma 3.1, represents a relative deviation of $\tilde{V}$ from V. The shifted BFGS update $A_+ = \gamma\tilde{V}UU^T\tilde{V}^T + \varrho\tilde{s}\tilde{s}^T/\tilde{b}$, see (3.5), multiplies columns of U by $\sqrt{\gamma}\,\tilde{V}$. In the BFGS update, see [9], which can be written in the form $H_+ = \gamma VHV^T + \varrho ss^T/b$, multiplication by $\sqrt{\gamma}\,V$ instead of $\sqrt{\gamma}\,\tilde{V}$ is performed. Thus if $A \approx H$ and $\|A\|$ is great compared to $\|\varrho\tilde{s}\tilde{s}^T/\tilde{b} - \varrho ss^T/b\|$ and if we want to have the shifted BFGS and the BFGS update not too different, ϕ should not be great.

When we chose μ close to unity in our numerical experiments, we often found a strongly dominant column of U (usually the first one), whose norm increased steadily. Denoting u the dominant column, $\bar{u} = (b/u^T y)u$ for $u^T y \neq 0$, we have $s \approx \xi u$ for some $\xi \in \mathcal{R}$ by (1.1), thus $s \approx \bar{u}$ and by (3.9) we get $|\tilde{V}u|/|u| = |\tilde{V}\bar{u}|/|\bar{u}| \approx \phi$. Therefore for $\sqrt{\gamma}\phi > 1$ we can expect exponential growth of the norm of this column and probably also convergence loss. We can reason similarly in case of a cluster of dominant linearly dependent columns of U. Setting $\sqrt{\gamma}\phi = 1$, we obtain $\mu_1 = 1/\left(1 + \sqrt{\gamma}\sqrt{1 - b^2/(\hat{a}|s|^2)}\right)$. This value can serve as a reasonable maximum of μ and should be multiplied by coefficient $\varepsilon > 0$ with the properties

- if $U^T y = 0$ then $\varepsilon = 1$ because $\tilde{V}U = U$ and it is not necessary to decrease μ,
- if $\bar{a} = |U^T y|^2 > 0$ then $\varepsilon < 1$ to moderate possible convergence loss.

The choice $\varepsilon = \sqrt{1 - \bar{a}/a} = \sqrt{\zeta \hat{a}/a}$ represents a simple possibility how to satisfy these conditions. Moreover, this value of μ effectively damps down a possible growth of $\|U\|$ - better than the scaling parameter γ in μ_1 above; for this reason we omit γ in μ_1. Multiplying μ_1 (without γ) by ε, we obtain finally

$$\mu = \frac{\sqrt{1 - \bar{a}/a}}{1 + \sqrt{1 - b^2/(\hat{a}|s|^2)}}. \tag{3.10}$$

This value of μ has the following interesting property.

Theorem 3.2. *Let $A = 0$. Then matrix $H_+ = \zeta_+ I + A_+$ with value (3.10), where A_+ is given by (3.4), is optimally conditioned.*

Proof. If $A = 0$, formula (3.4) (where we omit the last two terms) gives $H_+ = \zeta_+ I + \varrho\, \tilde{s}\tilde{s}^T/\tilde{b}$, which yields $H_+^{-1} = (1/\zeta_+)[I - \tilde{s}\tilde{s}^T/(\sigma\tilde{b} + |\tilde{s}|^2)]$ by (3.2). Thus $\|H_+\| = \varrho(\sigma + |\tilde{s}|^2/\tilde{b})$, $\|H_+^{-1}\| = 1/\zeta_+ = 1/(\varrho\sigma)$, $\kappa_+ = \|H_+\|\|H_+^{-1}\| = 1 + |\tilde{s}|^2/(\sigma\tilde{b})$. By (3.3), (3.7) and denoting again $\nu = \mu/(1-\mu)$, we obtain

$$\begin{aligned}
\kappa_+ &= 1 + \frac{\hat{a}}{\nu b^2}\left|\frac{s - \mu(b/\hat{a})y}{1-\mu}\right|^2 = 1 + \frac{\hat{a}}{\nu b^2}\left|s(1+\nu) - \nu\frac{b}{\hat{a}}y\right|^2 \\
&= 1 + \frac{\hat{a}}{\nu b^2}\Big(|s|^2(1+\nu)^2 - \frac{b^2}{\hat{a}}(\nu^2 + 2\nu)\Big) = 1 + \frac{\hat{a}}{\nu b^2}|s|^2 + (\nu+2)\Big(\frac{\hat{a}}{b^2}|s|^2 - 1\Big),
\end{aligned}$$

which gives the equation for the local minimum of function $\kappa_+(\nu)$

$$\frac{\hat{a}}{b^2}|s|^2\Big(1 - \frac{1}{\nu^2}\Big) = 1$$

with the positive root $\nu = 1/\sqrt{1 - b^2/(\hat{a}|s|^2)}$. By $\bar{a} = 0$, this leads to (3.10). $\square$

Formula (3.10) gives good results with update (3.4) without any corrections, with the exception of the first five to ten iterations, when it must be corrected, e.g. in the following way

$$\mu = \min\left(\max\left(\sqrt{1 - \bar{a}/a}\Big/\Big(1 + \sqrt{1 - b^2/(\hat{a}|s|^2)}\Big), 0.2\right), 0.8\right), \tag{3.11}$$

because our reasoning leading to (3.10) was simplified and the shifted VM methods effectivity is very sensitive to the shift parameter determination in the first iterations.

3.3 The shifted DFP method

If $A_1 = 0$ and $\gamma_k = \varrho_k = 1$, $k \ge 1$, then the shifted DFP method with

$$A_{k+1} = A_k + \frac{\tilde{s}_k\tilde{s}_k^T}{\tilde{b}_k} - \frac{A_k y_k y_k^T A_k}{\bar{a}_k}, \quad k \ge 1, \tag{3.12}$$

has an interesting property.

Theorem 3.3. *Consider the sequence of matrices A_k, $k \ge 1$, satisfying (3.12) with $A_1 = 0$ (if $k = 1$ we omit the last term) and $\bar{a}_k \ne 0$, $k \ge 2$. Then*

$$A_{k+1} = \frac{\tilde{s}_k \tilde{s}_k^T}{\tilde{b}_k}, \quad k \ge 1. \tag{3.13}$$

Proof (by induction). For $k = 1$, (3.13) holds by assumption. Suppose that (3.13) holds for index $k-1$. Then

$$A_k y_k = \frac{\tilde{s}_{k-1} \tilde{s}_{k-1}^T}{\tilde{b}_{k-1}} y_k = \frac{\tilde{s}_{k-1}^T y_k}{\tilde{b}_{k-1}} \tilde{s}_{k-1},$$

so

$$A_{k+1} = \frac{\tilde{s}_{k-1} \tilde{s}_{k-1}^T}{\tilde{b}_{k-1}} + \frac{\tilde{s}_k \tilde{s}_k^T}{\tilde{b}_k} - \frac{\tilde{b}_{k-1}}{(\tilde{s}_{k-1}^T y_k)^2} \frac{(\tilde{s}_{k-1}^T y_k)^2}{\tilde{b}_{k-1}^2} \tilde{s}_{k-1} \tilde{s}_{k-1}^T = \frac{\tilde{s}_k \tilde{s}_k^T}{\tilde{b}_k}$$

by (3.12), thus (3.13) is proved for index k. $\square$

Consider now that the line search is perfect, i.e. $s_k^T g_{k+1} = 0$, $k \ge 1$. Then

$$\tilde{s}_k^T g_{k+1} = s_k^T g_{k+1} - \zeta_{k+1} y_k^T g_{k+1} = -\zeta_{k+1} y_k^T g_{k+1},$$

$k \ge 1$, by (3.2) and (3.3). Thus using (1.1), (3.1) and (3.13), we can write for $k \ge 1$

$$d_{k+1} = -H_{k+1} g_{k+1} = -\zeta_{k+1} g_{k+1} - \frac{\tilde{s}_k \tilde{s}_k^T}{\tilde{b}_k} g_{k+1} = \zeta_{k+1} \Big(-g_{k+1} + \frac{y_k^T g_{k+1}}{\tilde{b}_k} \tilde{s}_k \Big). \tag{3.14}$$

We can interpret (3.14) as the shifted conjugate gradient method.

If $A_1 = 0$ is chosen, regardless of whether $\gamma_k = \varrho_k = 1$ holds, $k \ge 1$, the shifted DFP method for $\bar{a}_k \ne 0$, $k \ge 2$, always generates a sequence of matrices of rank at most one. This follows from $A_2 = (\varrho_1/\tilde{b}_1)\tilde{s}_1 \tilde{s}_1^T$ and from the product form of the shifted DFP method (3.6)) for $\eta = 0$, which shows that the rank of the updated matrix cannot increase. Therefore, this method does not accumulate information from previous iterations sufficiently, which probably causes its lower efficiency.

Very surprising results were obtained with the modified shifted DFP method which uses a modified quasi-Newton condition

$$A_+ y = \tilde{s} + \xi_1 A y,$$

where $0 < \xi_1 < \gamma$ (suitable values are $\xi_1 \le \gamma/2$). In this case, the update has the form

$$\frac{1}{\gamma} A_+ = A + \frac{\varrho}{\gamma} \frac{\tilde{s} \tilde{s}^T}{\tilde{b}} - \xi_2 A y y^T A, \quad \xi_2 = \frac{1}{\bar{a}} \Big(1 - \frac{\xi_1}{\gamma} \Big). \tag{3.15}$$

This method can be much more efficient than the standard shifted DFP method, as shown in Section 3.6.

The choice $\xi_1 = \gamma(1 - \bar{a}/a)$, i.e. $\xi_2 = 1/a$ is another interesting variant of the method.

3.4 Global convergence

In this section we use the following assumptions.

Assumption 3.1. *The objective function $f : \mathcal{R}^N \to \mathcal{R}$ is uniformly convex and has bounded second derivatives (i.e. $0 < \underline{G} \le \underline{\lambda}(G(x)) \le \overline{\lambda}(G(x)) \le \overline{G} < \infty$, $x \in \mathcal{R}^N$, where $\underline{\lambda}(G(x))$ and $\overline{\lambda}(G(x))$ are the lowest and the greatest eigenvalues of the Hessian matrix $G(x)$).*

Assumption 3.2. *Parameters ϱ_k and μ_k of the shifted VM method are uniformly positive and bounded (i.e. $0 < \underline{\varrho} \le \varrho_k \le \overline{\varrho} < \infty$, $0 < \underline{\mu} \le \mu_k \le \overline{\mu} < 1$, $k \ge 1$).*

Lemma 3.2. *Let $s \neq 0$, the objective function satisfy Assumption 3.1 and parameter μ satisfy Assumption 3.2. Then $y \neq 0$, $\tilde{s} \neq 0$, $b > 0$, $\tilde{b} > 0$, $\hat{a}/b \in [\underline{G}, \overline{G}]$ and $b/|s|^2 \ge \underline{G}$.*

Proof. Setting $G^I = \int_0^1 G(x + \xi s) d\xi$, one has $y = g_+ - g = G^I s$ and Assumption 3.1 gives $b = \int_0^1 s^T G(x + \xi s) s d\xi > 0$, which yields $y \neq 0$. Thus $\tilde{b} > 0$ by Assumption 3.2 and (3.7), which implies $\tilde{s} \neq 0$. Furthermore, setting $q = (G^I)^{1/2} s$, we obtain

$$\frac{\hat{a}}{b} = \frac{y^T y}{s^T y} = \frac{q^T G^I q}{q^T q} = \int_0^1 \frac{q^T G(x + \xi s) q}{q^T q} d\xi \in [\underline{G}, \overline{G}]$$

by Assumption 3.1. Similarly, $b/|s|^2 = s^T G^I s / s^T s = \int_0^1 s^T G(x + \xi s) s / s^T s \, d\xi \ge \underline{G}$. $\Box$

Theorem 3.4. *Let the objective function satisfy Assumption 3.1. Consider any shifted variable metric method satisfying (3.1)-(3.3) and Assumption 3.2, with the line search method fulfilling (1.1)-(1.2). If there is a constant $0 < C < \infty$ such that*

$$\mathrm{Tr} A_{k+1} \le \mathrm{Tr} A_k + C, \quad k \ge 1, \tag{3.16}$$

then (1.3) holds.

Proof. Since $\hat{a}/b \in [\underline{G}, \overline{G}]$ by Lemma 3.2, Assumption 3.2 implies $\zeta_{k+1} \in [\underline{\zeta}, \overline{\zeta}]$, $k \ge 1$, by (3.2) and (3.7), where $\underline{\zeta} = \underline{\varrho}\, \underline{\mu} \,/\, \overline{G}$ and $\overline{\zeta} = \overline{\varrho}\, \overline{\mu} / \underline{G}$. Using (3.16), one has

$$\|H_{k+1}\| \le \zeta_{k+1} + \|A_{k+1}\| \le \overline{\zeta} + \mathrm{Tr} A_{k+1} \le \overline{\zeta} + \mathrm{Tr} A_1 + C\, k \le \tilde{C}\,(k+1), \;\; k \ge 1,$$

where $\tilde{C} = \max(\overline{\zeta} + \mathrm{Tr} A_1, C)$. By (3.1), this gives

$$cos^2\theta_k \triangleq \frac{(g_k^T H_k g_k)^2}{g_k^T g_k \, g_k^T H_k^2 g_k} = \frac{g_k^T H_k g_k}{g_k^T g_k} \frac{g_k^T H_k g_k}{g_k^T H_k^2 g_k} \ge \zeta_k \frac{1}{\|H_k\|} \ge \frac{\underline{\zeta}}{\tilde{C}\, k}, \quad k \ge 1.$$

Thus $\sum_{k=1}^{\infty} cos^2\theta_k = \infty$ and (1.3) follows from Theorem 1.1. $\Box$

Theorem 3.5. *Let the objective function satisfy Assumption 3.1. Consider the shifted variable metric method (3.4) satisfying Assumption 3.2 and $\gamma_k \le 1$, $k \ge 1$, with the line search method fulfilling (1.1)-(1.2). If there is a constant $C < \infty$ such that*

$$\eta_k \left| \frac{\bar{a}_k}{\tilde{b}_k} \tilde{s}_k - A_k y_k \right|^2 \le C \frac{\bar{a}_k}{\tilde{b}_k} |\tilde{s}_k|^2 + |A_k y_k|^2, \quad k \ge 1, \tag{3.17}$$

then (1.3) holds.

Proof. From (3.4) and (3.17) we obtain (if $\bar a = 0$ we omit the two terms containing $\bar a$)

$$\frac{1}{\gamma}\mathrm{Tr}A_+ = \mathrm{Tr}A + \frac{\varrho}{\gamma}\frac{1}{\tilde b}|\tilde s|^2 - \frac{1}{\bar a}|Ay|^2 + \frac{\eta}{\bar a}\left|\frac{\bar a}{\tilde b}\tilde s - Ay\right|^2 \le \mathrm{Tr}A + \frac{\varrho}{\gamma}\frac{1}{\tilde b}|\tilde s|^2 + \frac{C}{\tilde b}|\tilde s|^2.$$

Since $|\tilde s|^2 = |s|^2 - \mu(2-\mu)b^2/\hat a \le |s|^2$ by (3.3), (3.7) and Assumption 3.2, we have $\tilde b/|\tilde s|^2 \ge (1-\mu)b/|s|^2 \ge (1-\overline{\mu})\underline{G}$ by (3.7), Assumption 3.2 and Lemma 3.2. Using inequality $\gamma \le 1$, we obtain

$$\mathrm{Tr}A_+ \le \gamma\mathrm{Tr}A + \frac{\varrho}{\tilde b}|\tilde s|^2 + \gamma\frac{C}{\tilde b}|\tilde s|^2 \le \mathrm{Tr}A + (\varrho + C)\frac{|\tilde s|^2}{\tilde b} \le \mathrm{Tr}A + \frac{\overline{\varrho} + C}{(1-\overline{\mu})\underline{G}},$$

which implies (1.3) by Theorem 3.4. □

Theorem 3.5 forms a basis for the hybrid globally convergent shifted VM method. We choose a constant C and parameters η_k, $k \ge 1$, which satisfy (3.17). Note that we can always choose $\eta > 0$ (the choice of η is irrelevant for $\bar a = 0$). Since choice $\eta = 0$ satisfies (3.17), the shifted DFP method is globally convergent. Also the modified shifted DFP method with $\xi_2 > 0$ is globally convergent owing to Theorem 3.4 and $\tilde b/|\tilde s|^2 \ge (1-\overline{\mu})\underline{G}$ (see the proof of Theorem 3.5). In this connection, our numerical experiments show that these methods are less sensitive to the choice of parameter σ.

Formula (3.17) shows that the uniform boundedness of $\bar a/\tilde b$ is crucial for the global convergence. If $\bar a/\tilde b$ is bounded, we can choose C in such a way that $\bar a/\tilde b \le C$. Then

$$\frac{C(\bar a/\tilde b)|\tilde s|^2 + |Ay|^2}{|(\bar a/\tilde b)\tilde s - Ay|^2} \ge \frac{|(\bar a/\tilde b)\tilde s|^2 + |Ay|^2}{|(\bar a/\tilde b)\tilde s - Ay|^2} \ge \frac{|(\bar a/\tilde b)\tilde s|^2 + |Ay|^2}{2(|(\bar a/\tilde b)\tilde s|^2 + |Ay|^2)} = \frac{1}{2},$$

so a reasonable value of η can be used.

3.5 Conditions for the superlinear rate of convergence

Lemma 3.3. *Consider any shifted variable metric method satisfying (3.1)-(3.3) and Assumption 3.2, with the line search method fulfilling (1.1)-(1.2) and such that $t_k = 1$ whenever this value satisfies (1.2). Suppose that $x_k \to x^\star$, where $x^\star$ satisfies the second order sufficient conditions for the local minimum of f (i.e. $g(x^\star) = 0$ and $G(x^\star)$ is positive definite). If $|\sigma_k - \zeta_k| \to 0$ and*

$$\lim_{k\to\infty} |\tilde s_k - A_k y_k|/|y_k| = 0,$$

then an index $k_0 \ge 1$ exists such that $\alpha_k = 1$, $k \ge k_0$, and $x_k \to x^\star$ superlinearly.

Proof. If $x^\star$ satisfies the second order sufficient conditions for the local minimum of f, then Assumption 3.1 is fulfilled in a neighbourhood of $x^\star$. Let x_k, $k \ge k_0$ be sufficiently close to $x^\star$ so Assumption 3.1 is satisfied. Then $\zeta_k \ge \underline{\zeta}$ (see proof of Theorem 3.4) and

$$|B_k s_k - y_k| \le \|B_k\|\,|s_k - H_k y_k| \le |s_k - H_k y_k|/\underline{\zeta} = |\tilde s_k + (\sigma_k - \zeta_k)y_k - A_k y_k|/\underline{\zeta}$$

by (3.1) and (3.3). Since $|y_k| = |\int_0^1 G(x_k + \xi s_k)s_k d\xi| \le \overline{G}|s_k|$, we obtain

$$\frac{\underline{\zeta}|B_k s_k - y_k|}{\overline{G}|s_k|} \le \frac{|\tilde s_k + (\sigma_k - \zeta_k)y_k - A_k y_k|}{|y_k|} \le \frac{|\tilde s_k - A_k y_k|}{|y_k|} + |\sigma_k - \zeta_k| \to 0$$

and we can use Theorem 1.2 with condition (1.7). □

Theorem 3.6. *Let the assumptions of Lemma 3.3 be satisfied. Consider the shifted variable metric method (3.4) with* $\gamma_k = \varrho_k = 1$ *and* $\eta_k \geq \underline{\eta} > 0$. *If*

$$\left(1 - \eta_k + \eta_k \frac{\bar{a}_k}{\tilde{b}_k}\right)\left(\frac{\bar{a}_k}{\tilde{b}_k}|\tilde{s}_k|^2 - |A_k y_k|^2\right) \geq 0 \tag{3.18}$$

and $\mathrm{Tr}A_k \leq C$, $k \geq 1$, *for some* $C < \infty$, *then the shifted variable metric method converges to* $x^\star$ *superlinearly.*

Proof. Using (3.4) with $\gamma = \varrho = 1$, we can write (if $\bar{a} = 0$ we omit the last two terms)

$$\mathrm{Tr}A_+ - \mathrm{Tr}A = \frac{|\tilde{s}|^2}{\tilde{b}} - \frac{|Ay|^2}{\bar{a}} + \frac{\eta}{\bar{a}}\left|\frac{\bar{a}}{\tilde{b}}\tilde{s} - Ay\right|^2,$$

which can be rewritten in the form

$$\mathrm{Tr}A_+ - \mathrm{Tr}A = \eta\frac{\hat{a}}{\tilde{b}}\left(\frac{|\tilde{s} - Ay|}{|y|}\right)^2 + \frac{1}{\bar{a}}\left(1 - \eta + \eta\frac{\bar{a}}{\tilde{b}}\right)\left(\frac{\bar{a}}{\tilde{b}}|\tilde{s}|^2 - |Ay|^2\right)$$

for $\bar{a} \neq 0$, or $\mathrm{Tr}A_+ - \mathrm{Tr}A = (|\tilde{s} - Ay|/|y|)^2\, \hat{a}/\tilde{b}$ otherwise. Now application of Lemma 3.3 and assumption (3.18) completes the proof if we realize that $\eta\hat{a}/\tilde{b} > \eta\hat{a}/b \geq \underline{\eta}\underline{G}$ by (3.7), Assumption 3.2 and Lemma 3.2 and that boundedness of $\{\mathrm{Tr}A_k\}$ together with $\mathrm{Tr}A_{k+1} - \mathrm{Tr}A_k \geq 0$, $k \geq 1$ imply $\mathrm{Tr}A_{k+1} - \mathrm{Tr}A_k \to 0$. □

The conditions used in Theorem 3.6 are relatively strong. First, we require $|\sigma_k - \zeta_k| = |\zeta_{k+1} - \zeta_k| \to 0$, which can be achieved by a suitable shift parameter determination strategy. Secondly, matrices A_k have to be bounded. This condition is not necessary for the global convergence. Third, the inequality (3.18) should hold.

Theorem 3.7. *Consider the shifted variable metric method (3.4) with* $\eta(\tilde{b} - \bar{a}) \leq \tilde{b}$ *(e.g.* $\eta = 1$*) and set* $\alpha = \hat{a}|s|^2/b^2$ *and* $\beta = \hat{a}|Ay|^2/(2\bar{a}b)$. *If* $\bar{a} = 0$ *or* $\beta^2 \leq \alpha - 1$ *then (3.18) holds for any* $\mu \in (0, 1)$. *Otherwise, if*

$$1 + \sqrt{\beta^2 - \alpha + 1} - \beta \leq \mu < 1 \quad \textit{or} \quad 0 < \mu \leq 1 - \sqrt{\beta^2 - \alpha + 1} - \beta, \tag{3.19}$$

then (3.18) holds.

Proof. Since assumption $\eta(\tilde{b} - \bar{a}) \leq \tilde{b}$ implies $1 - \eta + \eta\bar{a}/\tilde{b} \geq 0$, it suffices to examine the inequality $|Ay|^2 \leq |\tilde{s}|^2\, \bar{a}/\tilde{b}$ for $\bar{a} \neq 0$. By (3.7) and (3.3) we have $\tilde{b} = b(1 - \mu)$ and

$$|\tilde{s}|^2 = \left|s - \left(\mu\frac{b}{\hat{a}}\right)y\right|^2 = |s|^2 - 2\mu\frac{b^2}{\hat{a}} + \mu^2\frac{b^2}{\hat{a}} = \frac{b^2}{\hat{a}}(\alpha - 2\mu + \mu^2).$$

Using these relations, we can write condition $|Ay|^2 \leq |\tilde{s}|^2\, \bar{a}/\tilde{b}$ as the following quadratic inequality

$$\mu^2 - 2\mu(1 - \beta) + \alpha - 2\beta \geq 0,$$

which is satisfied if the discriminant is negative, i.e. $1 + \beta^2 - \alpha < 0$, or if (3.19) holds. □

Note that $\alpha \geq 1$ always holds by the Schwartz inequality and that for $\alpha > 1$ (which occurs when vectors s, y are linearly independent) we can always find $\mu < 1$ satisfying the first inequality in (3.19).

It is very difficult to utilize the above conditions in general. One reason is that (3.19) frequently gives values close to unit, which are unsuitable. Therefore our conditions for the superlinear rate of convergence can conflict with the numerical stability or with conditions for the global convergence. Nevertheless, the superlinear rate of convergence appears in some cases. We have investigated this phenomenon (i.e. condition (1.7)) numerically and found that approximately 10% of cases indicate such behaviour. Moreover, the following computational experiments show a surprisingly good efficiency of the shifted VM methods.

3.6 Computational experiments

The shifted VM methods were tested using a collection of 92 relatively difficult problems with optional dimension chosen from [10], [12] and [15] (problems given in [10] can be downloaded from `http://www.cs.cas.cz/~luksan/test.html`). We have used the dimension $n = 50$ and the final precision $|g(x^\star)| \leq 10^{-6}$. The results of our experiments are given in three tables, where NIT is the total number of iterations (over all 92 problems), NFV the total number of function evaluations and NRS the total number of restarts. 'Fail' denotes the number of problems which were not solved successfully (usually NFV reached its limit). 'Ratio' denotes the number of iterations with $\eta < 1$ (for hybrid strategy with controlled η). We chose $\varrho = \gamma = 1$ for shifted VM methods.

The first row of Table 1 gives results for the shifted BFGS method with choice (3.10) of the shift parameter μ and corrections (3.11) in the first six iterations; this choice is also used in the next tables. The next six rows demonstrate an influence of the constant parameter μ on the efficiency of the shifted BFGS method (the value 0.22 is in range (3.8)). We see that the convergence is lost when $\mu \geq 1/2$. The last four rows contain results for the standard BFGS method with various scaling strategies: 1 – scaling suppressed, 2 – preliminary scaling (see [16]), 3 – interval scaling (see [9]), 4 – controlled scaling (see [8]). This table demonstrates the high efficiency of the shifted BFGS method. It is much more efficient than the standard BFGS method with usually used preliminary scaling. Better results were obtained only by using the standard BFGS method with interval and controlled scaling. However, the convergence theory is not yet developed for these scaling strategies (see e.g. [14]).

Table 2 gives results for various choices of parameter η ($\eta = 0$ corresponds to the shifted DFP method). We can see that the shifted DFP method is rather inefficient (but better than the unscaled standard DFP method (DFP/1 in Table 3). The shifted BFGS method is very efficient, although global convergence was not proved for it in this paper. But experiments with the hybrid strategy described in Section 3.4 (in the last three rows, C is the constant in (3.17)) show that value $\eta < 1$ appears rarely (for $C = 10$ only in 1.12% cases). This fact shows that the shifted BFGS method is very robust and reliable for practical computations.

The first five rows in Table 3 contain results for the modified shifted DFP method ($\eta = 0$) with various choices of the relaxation parameters ξ_1 or ξ_2 (see (3.15)). The last four rows contain results for the standard DFP method with various scaling strategies: 1 – scaling suppressed, 2 – preliminary scaling, 3 – interval scaling, 4 – controlled scaling. This table demonstrates that a reasonable choice of relaxation parameters (e.g. $\xi_1 = 0.3$ or $\xi_2 = 1/a$) highly increases efficiency of the shifted DFP method. Moreover, the shifted DFP method is much more efficient than the standard DFP method with usually used preliminary scaling. Better results were obtained only by using the standard DFP method with interval and controlled scaling.

Method	NIT	NFV	NRS	Fail
SBFGS	11256	12178	1	-
$\mu = 0.22$	12252	13992	6	-
$\mu = 0.32$	12277	15093	5	-
$\mu = 0.42$	12966	18429	4	2
$\mu = 0.48$	16044	28357	6	3
$\mu = 0.50$	31388	65080	5	22
$\mu = 0.52$	24669	103575	49	44
BFGS/1	14075	22238	14	2
BFGS/2	14939	16335	3	1
BFGS/3	9731	10963	2	-
BFGS/4	7912	9322	2	-

Table 1

Method	NIT	NFV	NRS	Fail	Ratio (%)
$\eta = 0.0$	46010	48237	92	8	
$\eta = 0.5$	13262	14096	3	-	
$\eta = 1.0$	11256	12178	1	-	
$\eta = 1.5$	11117	12410	5	-	
$\eta = 2.0$	11403	13137	5	1	
$C = 0$	12412	13383	2	-	11.61
$C = 2$	11612	12570	2	-	2.77
$C = 10$	11373	12310	2	-	1.12

Table 2

Method	NIT	NFV	NRS	Fail
SDFP : $\xi_1 = 0.0$	46010	48237	92	8
SDFP : $\xi_1 = 0.1$	18707	19844	12	-
SDFP : $\xi_1 = 0.3$	15360	16726	7	1
SDFP : $\xi_1 = 0.5$	16315	19244	9	1
SDFP : $\xi_2 = 1/a$	15393	16189	5	-
DFP/1	79464	83895	6	35
DFP/2	94608	96309	4	42
DFP/3	11836	15196	4	1
DFP/4	11836	14884	1	1

Table 3

4 Limited-memory methods

All methods investigated in this section belong to shifted VM methods; they satisfy (3.1)-(3.3) and (3.7) with (positive semidefinite) matrix $A_k = U_k U_k^T$, where U_k, $k \ge 1$, is a rectangular matrix, and use the VM update

$$A_{k+1} = \gamma_k V_k A_k V_k^T, \tag{4.1}$$

$k \ge 1$, where V_k has the form $I + p_k q_k^T$ for the type 1 methods, or $I + p_1^k y_k^T + p_2^k s_k^T B_k$, where $B_k = H_k^{-1}$, for the type 2 methods. Thus we need to store only matrix U_k, which can be updated using relation

$$U_{k+1} = \sqrt{\gamma_k}\, V_k U_k, \tag{4.2}$$

$k \ge 1$. In the subsequent analysis we use the following notation

$$\begin{array}{llll} a_k = y_k^T H_k y_k, & b_k = s_k^T y_k, & c_k = s_k^T B_k s_k, & \delta_k = a_k c_k - b_k^2, \\ \bar{a}_k = y_k^T A_k y_k, & \bar{b}_k = s_k^T B_k A_k y_k, & \bar{c}_k = s_k^T B_k A_k B_k s_k, & \bar{\delta}_k = \bar{a}_k \bar{c}_k - \bar{b}_k^2, \\ \hat{a}_k = y_k^T y_k, & \hat{b}_k = s_k^T B_k y_k, & & \end{array}$$

$k \ge 1$. Note that the Schwartz inequality implies $\delta_k \ge 0$ and $\bar{\delta}_k \ge 0$. To simplify the notation we again frequently omit index k, replace index $k+1$ by symbol $+$ and consider also non-unit values of γ_k and ϱ_k in subsequent analysis as it is usual in case of VM methods (see [9]).

The shifted VM methods presented in Section 3, particularly in the quasi-product form (3.5), are ideal as starting methods. Setting $U_+ = \left(\sqrt{\varrho/\tilde{b}}\, \tilde{s}\right)$ in the first iteration, every update (3.5) modifies U and adds one column $\sqrt{\varrho/\tilde{b}}\, \tilde{s}$ to U_+. Thus in this section we will assume that the starting iterations have been executed and that matrix U has $m \ge 1$ columns in all iterations.

The type 1 methods are simpler and have many interesting properties, but the type 2 methods appear to be more efficient in practice. Note that the shifted DFP method (see Section 3.3) can be an example of the type 1 method.

4.1 Type 1 methods

Setting $V = I + pq^T$ in (4.1) one has

$$(1/\gamma)A_+ = A + Aqp^T + pq^TA + (q^TAq)pp^T. \tag{4.3}$$

Denoting $\tau = p^Ty$, quasi-Newton condition (3.2) gives

$$\begin{aligned} w &\triangleq (\varrho/\gamma)\tilde{s} - Ay = \tau Aq + (q^TAy + \tau q^TAq)p, & (4.4)\\ w^Ty &= (\varrho/\gamma)\tilde{b} - \bar{a} = \tau^2 q^TAq + 2\tau q^TAy. & (4.5) \end{aligned}$$

From (4.5) we obtain $(q^TAy + \tau q^TAq)^2 = (q^TAy)^2 + q^TAq\; w^Ty$ after rearrangement. Denoting $D = q^TAy + \tau q^TAq$, we have

$$D = q^TAy + \tau q^TAq, \qquad D^2 = (q^TAy)^2 + q^TAq\; w^Ty. \tag{4.6}$$

Thus we can calculate vector p for given q using formulas

$$\tau = (D - q^TAy)/q^TAq, \qquad p = (w - \tau Aq)/D \tag{4.7}$$

by (4.4) (first we calculate D^2, then τ and p). Since $D^2 > 0$ must hold, (4.6), (4.5) and (3.7) give the conditions (the inequality right side can be negative)

$$q^T A q \neq 0, \qquad \mu < 1 - \frac{\gamma}{\varrho b}\left(\bar{a} - (q^T A y)^2 / q^T A q\right). \tag{4.8}$$

Note that $\bar{a} \geq (q^T A y)^2 / q^T A q$ by the Schwartz inequality.

General type 1 method expression

Expression (4.3) can be written in another form. By (4.6) and (4.7) one has

$$\frac{1}{\gamma} A_+ - A = (q^T A y + \tau q^T A q) \frac{Aq(w - \tau Aq)^T + (w - \tau Aq) q^T A}{D^2} + q^T A q \frac{(w - \tau Aq)(w - \tau Aq)^T}{D^2}.$$

Rearranging this and using (4.5), we obtain the following formulas

$$\begin{aligned} \frac{1}{\gamma} A_+ - A &= \frac{q^T A q \; w w^T + q^T A y \left(A q w^T + w q^T A\right) - w^T y \; A q q^T A}{D^2} \\ &= \frac{w w^T}{w^T y} - \frac{w^T y}{D^2}\left(I - \frac{w y^T}{w^T y}\right) A q q^T A \left(I - \frac{y w^T}{w^T y}\right) \\ &= \frac{q^T A q}{D^2}\left(w + \frac{q^T A y}{q^T A q} A q\right)\left(w + \frac{q^T A y}{q^T A q} A q\right)^T - \frac{A q q^T A}{q^T A q}. \end{aligned} \tag{4.9}$$

In order to obtain the form closer to (3.4), the term $w w^T / w^T y$ (which is $(1/\gamma) A_+ - A$ for the shifted rank-one update, by analogy with the Broyden class, see [3]) in the second formula can be written in the following way:

$$\frac{w w^T}{w^T y} = \frac{\varrho}{\gamma} \frac{\tilde{s} \tilde{s}^T}{\tilde{b}} - \frac{A y y^T A}{\bar{a}} + \frac{\varrho / \gamma}{\bar{a}(\varrho/\gamma - \bar{a}/\tilde{b})}\left(\frac{\bar{a}}{\tilde{b}} \tilde{s} - A y\right)\left(\frac{\bar{a}}{\tilde{b}} \tilde{s} - A y\right)^T,$$

from which e.g. the following forms of the shifted BFGS formula can be derived

$$\frac{1}{\gamma} A_+ - A = \frac{w w^T}{w^T y} - \frac{1}{w^T y}\left(\frac{\bar{a}}{\tilde{b}} \tilde{s} - A y\right)\left(\frac{\bar{a}}{\tilde{b}} \tilde{s} - A y\right)^T = \frac{w w^T}{w^T y} - \frac{w^T y}{\tilde{b}^2}\left(I - \frac{w y^T}{w^T y}\right) \tilde{s} \tilde{s}^T \left(I - \frac{y w^T}{w^T y}\right).$$

Note that we need not know vector q for updating. All relations can be based on the vector $\tilde{q} = U^T q \in \mathcal{R}^m$. In that case we use only update (4.2), in the form

$$U_+ = \sqrt{\gamma}(U + p \tilde{q}^T),$$

and rewrite the relations containing q in corresponding way, e.g. $D^2 = (y^T U \tilde{q})^2 + |\tilde{q}|^2 w^T y$.

Choice of vector parameter q

Effectivity of type 1 methods is considerably dependent on the choice of vector q. Good results were obtained only for $q = Bs$ and $q = y$. Since scaling of q has no influence on update (4.3), we choose $q = Bs + \vartheta y$, $\vartheta \in \mathcal{R}$. Then $Aq = ABs + \vartheta Ay = s - \zeta Bs + \vartheta Ay$ by (3.1) and

$$q^T A y = \bar{b} + \vartheta \bar{a}, \quad q^T A q = \bar{c} + 2\vartheta \bar{b} + \vartheta^2 \bar{a}, \quad D^2 = (\bar{c} + 2\vartheta \bar{b} + \vartheta^2 \bar{a}) \tilde{b} \varrho / \gamma - \bar{\delta} \tag{4.10}$$

by (4.6). The second condition in (4.8) has now the form (the right side can be negative)

$$\mu < 1 - \frac{\gamma}{\varrho b}\left(\bar{\delta}/q^T A q\right). \tag{4.11}$$

A suitable value of ϑ can be obtained by comparison between the Broyden class (see [3], [9]) and expression (4.9), where we set $\zeta = \sigma = 0$ (in that case relation (4.9) represents the Broyden update with A and $\tilde{s}$ replaced by H and s). In view of fulfilling the quasi-Newton condition, it suffices to compare only one term, e.g. containing $Hyy^T H$. The corresponding coefficient is $(\eta - 1)/a$ for the Broyden class and $(q^T A q - 2\vartheta q^T A y - \vartheta^2 w^T y)/D^2$ for (4.9). Using (4.10) and (4.5), we obtain

$$\frac{\eta - 1}{a} = \frac{q^T A q - 2\vartheta q^T A y - \vartheta^2 w^T y}{D^2} = \frac{c - \vartheta^2 b\varrho/\gamma}{(c + 2\vartheta b + \vartheta^2 a)b\varrho/\gamma - \delta}, \tag{4.12}$$

which can be rearranged in the form

$$\frac{\eta}{b} = \frac{b + (c + 2\vartheta b)\varrho/\gamma}{(c + 2\vartheta b + \vartheta^2 a)b\varrho/\gamma - \delta}. \tag{4.13}$$

For the BFGS ($\eta = 1$) and the DFP ($\eta = 0$) updates, relations (4.12) and (4.13) give

$$\vartheta^{BFGS} = \pm\sqrt{\frac{\gamma}{\varrho}\frac{c}{b}}, \qquad \vartheta^{DFP} = -\frac{1}{2}\left(\frac{\gamma}{\varrho} + \frac{c}{b}\right). \tag{4.14}$$

It is interesting that the positive value of ϑ^{BFGS} gives very good results, while the negative one is not suitable for type 1 methods. To calculate ϑ from (4.12)-(4.13), it is useful to set

$$\chi = (b/a)\,(1 - \eta)/\eta.$$

Dividing (4.12) by (4.13), we obtain

$$\chi = \frac{\vartheta^2 b\varrho/\gamma - c}{b + (c + 2\vartheta b)\varrho/\gamma}.$$

This gives the quadratic equation $\vartheta^2 - 2\vartheta\chi - \chi\,(\gamma/\varrho + c/b) - (\gamma/\varrho)c/b = 0$, which has the roots

$$\vartheta = \chi \pm \sqrt{(\chi + \gamma/\varrho)\,(\chi + c/b)} \tag{4.15}$$

for any χ, excepting the values inside the interval with limits $-\gamma/\varrho$, $-c/b$. Note that the choice $\chi = -\gamma/\varrho$ corresponds to the rank-one Broyden update, see [3].

An update based on the Broyden class

Any attempt to approximate the Broyden class is complicated for the type 1 methods. Thus we present only one method, motivated by the BFGS update, with $q = Bs + \vartheta y$. From (4.7) we get

$$p = [(\varrho/\gamma)\tilde{s} - (1 + \tau\vartheta)Ay - \tau ABs]/D. \tag{4.16}$$

Setting $\zeta = \sigma = 0$ as above, we convert A to H and (4.16) to $p_0 = \alpha s + \beta H y$, where $\alpha = (\varrho/\gamma - \tau)/D$, $\beta = -(1+\tau\vartheta)/D$. Since the term of (4.3) (after rearrangement, with A and p replaced by H and p_0), which contains $Hyy^T H$, has coefficient $2\vartheta\beta + \beta^2 q^T A q$ and corresponding coefficient is zero for the BFGS update, we choose $\beta = 0$, i.e. $1 + \tau\vartheta = 0$. By (4.7) and (4.10) one has

$$0 = 1 + \tau\vartheta = (q^T A q - \vartheta q^T A y + \vartheta D)/q^T A q = (\bar{c} + \vartheta \bar{b} + \vartheta D)/q^T A q,$$

which yields $D = -\bar{b} - \bar{c}/\vartheta$ and thus

$$D^2 + \bar{\delta} = (\bar{b} + \bar{c}/\vartheta)^2 + \bar{a}\bar{c} - \bar{b}^2 = (\bar{c} + 2\vartheta\bar{b} + \vartheta^2 \bar{a})\bar{c}/\vartheta^2.$$

Comparing this with (4.10), we get $\bar{c}/\vartheta^2 = \tilde{b}\varrho/\gamma$ and thus we can calculate vectors p, q for any $\mu \in (0, 1)$, using the formulas (we again choose only positive ϑ, see comments after (4.14))

$$\vartheta = \sqrt{\frac{\gamma\,\bar{c}}{\varrho\,\tilde{b}}}, \qquad q = Bs + \vartheta y, \qquad p = -\frac{(\varrho/\gamma)\vartheta\,\tilde{s} + s - \zeta B s}{\bar{c} + \vartheta \bar{b}} \tag{4.17}$$

by (4.16) and (3.1). Note that it is possible to have $\vartheta = \vartheta^{BFGS}$ (see (4.14)) simultaneously, by setting $\mu = 1 - \bar{c}/c$ due to (3.7); this choice is, however, not so efficient as (4.17) with suitable μ (e.g. given by (3.10)).

4.2 Type 2 methods

For the best known choice $q = Bs + \vartheta y$ one has $V = I + \vartheta p y^T + p s^T B$ for the type 1 methods. To have more free parameters, we investigate the case $V = I + p_1 y^T + p_2 s^T B$ in this section. From (4.1) we have

$$\frac{1}{\gamma} A_+ = A + p_1 y^T A + A y p_1^T + p_2 s^T B A + A B s p_2^T + \bar{a} p_1 p_1^T + \bar{b}(p_1 p_2^T + p_2 p_1^T) + \bar{c} p_2 p_2^T. \tag{4.18}$$

Denoting $\tau_1 = 1 + p_1^T y$, $\tau_2 = p_2^T y$, the quasi-Newton condition (3.2) gives

$$\begin{aligned} (\bar{a}\tau_1 + \bar{b}\tau_2)p_1 + (\bar{b}\tau_1 + \bar{c}\tau_2)p_2 + \tau_1 A y + \tau_2 A B s &= (\varrho/\gamma)\tilde{s}, & (4.19)\\ \bar{a}\tau_1^2 + 2\bar{b}\tau_1\tau_2 + \bar{c}\tau_2^2 &= (\varrho/\gamma)\tilde{b}. & (4.20) \end{aligned}$$

Since we still assume $\tilde{b} > 0$, inequality $\bar{\delta} \ge 0$ together with (4.20) imply that at least one of values $\bar{a}$, $\bar{c}$ must be nonzero. We will use the following notation

$$v_1 = \bar{c} A y - \bar{b} A B s, \qquad v_2 = \bar{a} A B s - \bar{b} A y, \qquad q_1 = \bar{\delta} p_1 + v_1, \qquad q_2 = \bar{\delta} p_2 + v_2$$

and identities $v_1^T y = \bar{\delta}$, $v_2^T y = 0$ and

$$q_i^T y = \bar{\delta}\tau_i,\ i = 1, 2, \qquad \bar{a}(v_1 v_1^T + \bar{\delta} A B s s^T B A) = \bar{c}(v_2 v_2^T + \bar{\delta} A y y^T A). \tag{4.21}$$

Lemma 4.1. *Let $\bar{\delta} = 0$. Then $v_1 = v_2 = q_1 = q_2 = 0$.*

Proof. Vectors Ay, ABs are proportional by assumption and the same proportionality is between $\bar{a}$, $\bar{b}$ and also between $\bar{b}$, $\bar{c}$, which gives the desired assertion. $\square$

General type 2 method expression

First we will suppose that $\bar{a} \neq 0$ and that vectors p_1 and p_2 are chosen such that $\bar{a}\tau_1 + \bar{b}\tau_2 \neq 0$ and denote $\tilde{p} = \bar{a}p_1 + \bar{b}p_2$. Our approach is based on the following result.

Lemma 4.2. *Let $\bar{a} \neq 0$ and $\omega_1 \triangleq \bar{a}\tau_1 + \bar{b}\tau_2 \neq 0$. Then*

$$\omega_1^2 = \bar{a}\tilde{b}\varrho/\gamma - \bar{\delta}\tau_2^2, \qquad q_2 q_2^T + \bar{\delta}(\tilde{p} + Ay)(\tilde{p} + Ay)^T = \bar{q}_2\bar{q}_2^T + \bar{a}\bar{\delta}(\varrho/\gamma)\tilde{s}\tilde{s}^T/\tilde{b},$$

where

$$\bar{q}_2 = \left(q_2 - (q_2^T y/\tilde{b})\tilde{s}\right)/(|\omega_1|\omega_2), \qquad \omega_2 = 1/\sqrt{\bar{a}\tilde{b}\varrho/\gamma}.$$

Proof. The first relation readily follows from (4.20). By (4.19) and (4.21) one has

$$\begin{aligned} \omega_1(\tilde{p} + Ay) &= (\bar{a}\tau_1 + \bar{b}\tau_2)(\bar{a}p_1 + \bar{b}p_2 + Ay) = \bar{a}\left((\varrho/\gamma)\tilde{s} - \tau_2(\bar{b}p_1 + \bar{c}p_2 + ABs)\right) \\ &+ \bar{b}\tau_2(\bar{a}p_1 + \bar{b}p_2 + Ay) = \bar{a}(\varrho/\gamma)\tilde{s} - \tau_2\bar{\delta}p_2 - \tau_2 v_2 = \bar{a}(\varrho/\gamma)\tilde{s} - \tau_2 q_2 \\ &= \bar{a}(\varrho/\gamma)\tilde{s} - \tau_2\left(|\omega_1|\omega_2\,\bar{q}_2 + (\bar{\delta}\tau_2/\tilde{b})\tilde{s}\right) = |\omega_1|(|\omega_1|\tilde{s}/\tilde{b} - \tau_2\omega_2\,\bar{q}_2), \end{aligned} \tag{4.22}$$

thus

$$\begin{aligned} &\bar{\delta}(\tilde{p} + Ay)(\tilde{p} + Ay)^T + q_2 q_2^T = \bar{\delta}(|\omega_1|\tilde{s}/\tilde{b} - \tau_2\omega_2\,\bar{q}_2)(|\omega_1|\tilde{s}/\tilde{b} - \tau_2\omega_2\,\bar{q}_2)^T \\ + \;& (\bar{\delta}\tau_2\tilde{s}/\tilde{b} + |\omega_1|\omega_2\,\bar{q}_2)(\bar{\delta}\tau_2\tilde{s}/\tilde{b} + |\omega_1|\omega_2\,\bar{q}_2)^T = \bar{a}\bar{\delta}(\varrho/\gamma)\tilde{s}\tilde{s}^T/\tilde{b} + \bar{q}_2\bar{q}_2^T. \end{aligned} \qquad \square$$

Before utilizing this lemma, we rewrite (4.18) in the following way

$$\begin{aligned} \bar{a}\left[(1/\gamma)A_+ - A\right] &= \tilde{p}y^T A + Ay\tilde{p}^T + p_2 v_2^T + v_2 p_2^T + \tilde{p}\tilde{p}^T + \bar{\delta}p_2 p_2^T \\ &= p_2 v_2^T + v_2 p_2^T + (\tilde{p} + Ay)(\tilde{p} + Ay)^T - Ayy^T A + \bar{\delta}p_2 p_2^T. \end{aligned} \tag{4.23}$$

Since $\bar{\delta}(p_2 v_2^T + v_2 p_2^T) + \bar{\delta}^2 p_2 p_2^T = q_2 q_2^T - v_2 v_2^T$, we can use Lemma 4.2 to obtain $\bar{a}\bar{\delta}\left[(1/\gamma)A_+ - A\right] = \bar{a}\bar{\delta}(\varrho/\gamma)\tilde{s}\tilde{s}^T/\tilde{b} - \bar{\delta}Ayy^T A + \bar{q}_2\bar{q}_2^T - v_2 v_2^T$. Since $\bar{q}_2 = q_2$ for $\tau_2 = 0$, we can assume (without any change of A_+) that $\tau_2 = 0$ is chosen, which satisfies the condition $\omega_1 \neq 0$ by (4.20), and the update formula can be written in the form

$$\frac{1}{\gamma}A_+ = A + \frac{\varrho}{\gamma}\frac{\tilde{s}\tilde{s}^T}{\tilde{b}} - \frac{Ayy^T A}{\bar{a}} + \frac{q_2 q_2^T - v_2 v_2^T}{\bar{a}\bar{\delta}}, \qquad q_2^T y = 0 \tag{4.24}$$

for $\bar{\delta} \neq 0$. If $\bar{\delta} = 0$, one has $v_2 = q_2 = \bar{q}_2 = 0$ by Lemma 4.1, thus $\tilde{p} + Ay = \omega_1\,\tilde{s}/\tilde{b}$ by (4.22) and from (4.23) we get $(1/\gamma)A_+ = A + (\varrho/\gamma)\tilde{s}\tilde{s}^T/\tilde{b} - Ayy^T A/\bar{a}$ (which is the shifted DFP update, see Section 3) for any choice of p_2.

Proceeding similarly for $\bar{a} = 0$, thus $\bar{c} \neq 0$, we derive the following update formula

$$\frac{1}{\gamma}A_+ = A + \frac{\varrho}{\gamma}\frac{\tilde{s}\tilde{s}^T}{\tilde{b}} - \frac{ABss^T BA}{\bar{c}} + \frac{q_1 q_1^T - v_1 v_1^T}{\bar{c}\bar{\delta}}, \qquad q_1^T y = 0 \tag{4.25}$$

for $\bar{\delta} \neq 0$ and $(1/\gamma)A_+ = A + (\varrho/\gamma)\tilde{s}\tilde{s}^T/\tilde{b} - ABss^T BA/\bar{c}$ for $\bar{\delta} = 0$ (and any p_1); this update satisfies the shifted quasi-Newton condition by Lemma 4.1. Note that by

(4.21), update (4.25) can be written in the form (4.24) with $q_2 q_2^T/\bar{a}$ replaced by $q_1 q_1^T/\bar{c}$ for $\bar{a}\bar{c} \neq 0$, but then we can directly use (4.24).

To construct the type 2 update, we can proceed in the following way. If $\bar{\delta} \neq 0$ (thus also $\bar{a}\bar{c} \neq 0$ by $\bar{\delta} \geq 0$) we choose vector parameter q_2 satisfying $q_2^T y = 0$, i.e. $\tau_2 = 0$. Then $\tau_1 = \pm\sqrt{(\varrho/\gamma)\tilde{b}/\bar{a}}$ holds by (4.20), and by (4.19) we can calculate p_1 and p_2, using the formulas

$$p_2 = \frac{q_2 - v_2}{\bar{\delta}}, \qquad p_1 = \frac{1}{\bar{a}}\left(\sqrt{\frac{\varrho}{\gamma}\frac{\bar{a}}{\tilde{b}}}\tilde{s} - Ay - \bar{b}p_2\right). \tag{4.26}$$

Otherwise, if $\bar{\delta} = 0$ and $\bar{a} \neq 0$, we have found above that the update (the shifted DFP update) is independent of vector p_2. Thus we choose $p_2 = 0$ and calculate the corresponding p_1, using (4.26). Similarly, if $\bar{\delta} = 0$ and $\bar{a} = 0$, thus $\bar{c} \neq 0$ and $\bar{b} = 0$, the update is independent of vector p_1 and we choose $p_1 = 0$. Then $\tau_1 = 1$ and $\tau_2 = \pm\sqrt{(\varrho/\gamma)\tilde{b}/\bar{c}}$ holds by (4.20) and by (4.19) we can calculate p_2, using the formula

$$p_2 = \left(\sqrt{(\varrho/\gamma)\bar{c}/\tilde{b}}\;\tilde{s} - \sqrt{(\gamma/\varrho)\bar{c}/\tilde{b}}\,Ay - ABs\right)\Big/\bar{c}. \tag{4.27}$$

In case $\bar{\delta} = 0$, the choice of q_2 (or q_1) is irrelevant; therefore in this section we will suppose from now on that $\bar{\delta} \neq 0$, thus $\bar{a}\bar{c} \neq 0$.

A simple method based on the Broyden class

Comparing (4.24) with (3.4) for $\zeta = \sigma = 0$ (or with (4.44)) and denoting $z = as - bHy$, we get $(q_2 q_2^T - zz^T)/\delta = \eta\, zz^T/b^2$, which yields $q_2 = \pm\sqrt{1 + \eta\delta/b^2}\, z$ and thus

$$p_2 = \frac{\eta}{b(b \pm \sqrt{b^2 + \eta\delta})}\, z \tag{4.28}$$

by (4.26). For the BFGS update ($\eta = 1$) we have $p_2 = z/(b^2 \pm b\sqrt{ac})$. The described method consists in choice (4.28) of vector parameter p_2 with p_1 given by (4.26). Note that only the case with the minus sign is suitable here.

This method gives good results when also p_1 is linearly dependent only on s, Hy, i.e. when $\zeta = \sigma\sqrt{(\varrho/\gamma)\bar{a}/\tilde{b}}$ by (4.26). This yields the quadratic equation $\mu^2(\varrho/\gamma)\bar{a}b/(a - \bar{a})^2 + \mu - 1 = 0$, which has one positive root

$$\mu = 2\Big/\left(1 + \sqrt{1 + 4(\varrho/\gamma)\bar{a}b/(a - \bar{a})^2}\right). \tag{4.29}$$

For this value of μ and for $\eta = 1$ we obtain the following formulas by (4.26)

$$p_2 = \frac{as - bHy}{b(b - \sqrt{ac})}, \qquad p_1 = \frac{a - \bar{a}}{\mu\bar{a}b}s - \frac{Hy + \bar{b}p_2}{\bar{a}}. \tag{4.30}$$

A method with direction vector derived from the shifted Broyden class

Since $d_+ = -H_+ g_+ = -H_+ y - H_+ g = -\varrho s + H_+ Bd$ by (1.1), (3.2) and (3.3), it suffices to compare value $(1/\gamma)H_+ Bs$, which is

$$\sigma\frac{\varrho}{\gamma}Bs + \frac{\varrho}{\gamma}\frac{\tilde{s}^T Bs}{\tilde{b}}\tilde{s} + \frac{q_2^T Bs}{\bar{a}\bar{\delta}}q_2 \tag{4.31}$$

by $v_2^T Bs = \bar{\delta}$ for update (4.24) and

$$\sigma \frac{\varrho}{\gamma} Bs + \frac{\varrho}{\gamma} \frac{\tilde{s}^T Bs}{\tilde{b}} \tilde{s} + \frac{1}{\bar{a}} v_2 + \frac{\eta}{\bar{a}} \left(\frac{\bar{a}}{\tilde{b}} \tilde{s}^T Bs - \bar{b} \right) \left(\frac{\bar{a}}{\tilde{b}} \tilde{s} - Ay \right) \tag{4.32}$$

for update (3.4). Comparing (4.31) with (4.32), we obtain

$$\frac{q_2^T Bs}{\bar{\delta}} q_2 = v_2 + \eta \left(\frac{\bar{a}}{\tilde{b}} \tilde{s}^T Bs - \bar{b} \right) \left(\frac{\bar{a}}{\tilde{b}} \tilde{s} - Ay \right), \tag{4.33}$$

which implies

$$\frac{q_2^T Bs}{\bar{\delta}} = \pm \sqrt{1 + \frac{\eta}{\bar{\delta}} \left(\frac{\bar{a}}{\tilde{b}} \tilde{s}^T Bs - \bar{b} \right)^2}. \tag{4.34}$$

Combining (4.33) with (4.34), we can calculate q_2 for given η (obviously $q_2^T y = 0$) and then p_2 and p_1, using (4.26).

A method nearest to the shifted Broyden class

Denoting $\hat{w} = \sqrt{\eta \bar{\delta}} \left((\bar{a}/\tilde{b}) \tilde{s} - Ay \right)$ and comparing (4.24) with (3.4), we see that matrix $q_2 q_2^T$ should be as near as possible in some sense to matrix $M_2 = v_2 v_2^T + \hat{w} \hat{w}^T$. We will find q_2 satisfying the following problem

$$q_2 = \arg\min \{ \|M_2 - qq^T\|_F^2 : \ q \in \mathcal{R}^N \}, \qquad \text{s.t.} \ \ q^T y = 0 \tag{4.35}$$

(Frobenius matrix norm). Note that we also tried to minimize $\|M_2 - q_2 q_2^T\|^2$ (for $q_2 \in \text{span}\{v_2, \hat{w}\}$), but this was much more complicated and the results were not better.

To solve this problem, we need the following two lemmas.

Lemma 4.3. *Let M be symmetric. Consider the problem*

$$\bar{r} = \arg\min \{ \|M - rr^T\|_F^2 : \ r \in \mathcal{R}^N \}, \qquad \text{s.t.} \ \ r^T y = 0.$$

If $My = 0$ then $\bar{r}$ is the eigenvector of M, corresponding to the largest eigenvalue of M, with the norm equal to square root of this eigenvalue.

Proof. Define Lagrangian function

$$\mathcal{L}(r, \nu) = \frac{1}{4} \left\| M - rr^T \right\|_F^2 + \nu \, r^T y = \frac{1}{4} \left(\|M\|_F^2 - 2 \, r^T M r + |r|^4 \right) + \nu \, r^T y. \tag{4.36}$$

A local minimizer $\bar{r}$ satisfies the equation

$$\frac{\partial \mathcal{L}}{\partial r} = |r|^2 r - Mr + \nu \, y = 0, \tag{4.37}$$

which gives $\nu = (r^T My - |r|^2 r^T y)/\hat{a} = 0$ by assumption, thus $Mr = |r|^2 r$ by (4.37). From (4.36) we obtain

$$\mathcal{L}(r, \nu) = (\|M\|_F^2 - |r|^4)/4, \tag{4.38}$$

therefore eigenvector r should correspond to the largest eigenvalue equal to $|r|^2$. $\square$

Lemma 4.4. *The nonzero eigenvalues of matrix $M = uu^T + vv^T$ have the form*

$$\lambda = (|u|^2 + |v|^2)/2 \pm \sqrt{(|u|^2 - |v|^2)^2/4 + (u^T v)^2}.$$

If $u^T v = 0$, then $Mu = |u|^2 u$, $Mv = |v|^2 v$. Otherwise, if $u^T v \neq 0$, then the eigenvector corresponding to the largest eigenvalue λ_1 of M can be written in the form

$$(u^T v)u + (\lambda_1 - |u|^2)v \qquad \text{or} \qquad (\lambda_1 - |v|^2)u + (u^T v)v. \tag{4.39}$$

Proof. Denoting by r the eigenvector corresponding to the nonzero eigenvalue λ, we have

$$(u^T r)u + (v^T r)v = \lambda r. \tag{4.40}$$

Multiplying this by u, v, we obtain the system

$$\begin{array}{lcll} u^T r(|u|^2 - \lambda) & + & v^T r(u^T v) & = & 0, \\ u^T r(u^T v) & + & v^T r(|v|^2 - \lambda) & = & 0. \end{array} \tag{4.41}$$

Determinant of this system is zero, since at least one of values $u^T r$, $v^T r$ must be nonzero by (4.40) and $\lambda \neq 0$. This leads to equation $\lambda^2 - \lambda(|u|^2 + |v|^2) + |u|^2|v|^2 - (u^T v)^2 = 0$ with roots

$$(|u|^2 + |v|^2)/2 \pm \sqrt{(|u|^2 + |v|^2)^2/4 + (u^T v)^2 - |u|^2|v|^2},$$

which can be rearranged to the desired form. The rest readily follows from (4.40) and (4.41). □

Now we turn back to problem (4.35). The two largest eigenvalues of M_2 are

$$\lambda_{1,2} = (|v_2|^2 + |\hat{w}|^2)/2 \pm \sqrt{(|v_2|^2 - |\hat{w}|^2)^2/4 + (v_2^T \hat{w})^2}, \qquad \lambda_1 \geq \lambda_2 \geq 0 \tag{4.42}$$

by Lemma 4.4 and $q_2 = \sqrt{\lambda_1}\, q_0/|q_0|$ by Lemma 4.3, where the eigenvector q_0 of M_2 corresponding to λ_1 can be obtained by using Lemma 4.4. Note that it is better to calculate the first form in (4.39) for $|v| \geq |u|$ and the second one otherwise, because then the term, which we add to the square root term in the corresponding formula for $\lambda_1 - |u|^2$ or $\lambda_1 - |v|^2$, is positive.

Since $\|M_2 - q_2 q_2^T\|_F^2 = \|M_2\|_F^2 - |q_2|^4 = (\lambda_1^2 + \lambda_2^2) - \lambda_1^2 = \lambda_2^2$ by (4.38) and Lemma 4.3, we should choose parameters of update (3.4) in such a way to make λ_2 as small as possible, but the following theorem shows that the problem is more complicated.

Theorem 4.1. *Function $\lambda_2(\eta)$ is increasing for $\eta > 0$.*

Proof. Denote $\alpha = |\hat{w}|^2 - |v_2|^2$, $\beta = v_2^T \hat{w}$. Since $dv_2/d\eta = 0$, $d\hat{w}/d\eta = \hat{w}/(2\eta)$, $d|\hat{w}|^2/d\eta = |\hat{w}|^2/\eta$ and $d(v_2^T \hat{w})^2/d\eta = (v_2^T \hat{w})^2/\eta$, one has by Lemma 4.4 for $v_2^T \hat{w} \neq 0$

$$2\eta\lambda_2'(\eta) = |\hat{w}|^2 - \frac{(|\hat{w}|^2 - |v_2|^2)|\hat{w}|^2 + 2(v_2^T \hat{w})^2}{\sqrt{(|\hat{w}|^2 - |v_2|^2)^2 + 4(v_2^T \hat{w})^2}} = |\hat{w}|^2 - \frac{\alpha|\hat{w}|^2 + 2\beta^2}{\sqrt{\alpha^2 + 4\beta^2}} \tag{4.43}$$

and $\lambda_2' \geq 0$ when the numerator $\alpha|\hat{w}|^2 + 2\beta^2$ is negative or zero. Otherwise, we can equivalently multiply (4.43) by the positive number $|\hat{w}|^2 + (\alpha|\hat{w}|^2 + 2\beta^2)/\sqrt{\alpha^2 + 4\beta^2}$ to obtain on the right side

$$|\hat{w}|^4 - \frac{\alpha^2|\hat{w}|^4 + 4\alpha\beta^2|\hat{w}|^2 + 4\beta^4}{\alpha^2 + 4\beta^2} = 4\beta^2 \frac{|\hat{w}|^4 - \alpha|\hat{w}|^2 - \beta^2}{\alpha^2 + 4\beta^2} = 4\beta^2 \frac{|v_2|^2|\hat{w}|^2 - (v_2^T\hat{w})^2}{\alpha^2 + 4\beta^2} \geq 0$$

by the Schwartz inequality. It remains to prove the assertion in case $v_2^T\hat{w} = 0$. But then $\lambda_2 = |v_2|^2$ or $\lambda_2 = |\hat{w}|^2$ by Lemma 4.4 and again $\lambda_2' \geq 0$ holds. □

Since $\hat{w} = 0$ for $\eta = 0$, one has $\lambda_2(0) = 0$. Thus Theorem 4.1 shows that small positive values of η should be chosen - but not too small, because the shifted DFP method ($\eta = 0$) is not effective. In this situation, it is useful to know $\lambda_2'(0)$. Denoting $\bar{w} = \hat{w}/\sqrt{\eta}$, it readily follows from (4.43) that $\lambda_2'(0) = |\bar{w}|^2 - (v_2^T\bar{w})^2/|v_2|^2$ for $v_2 \neq 0$, $\lambda_2'(0) = 0$ otherwise. It follows from (4.42) that $\lambda_2'(0)$ is close to zero (i.e. vectors v_2, $\bar{w}$ are almost proportional) when e.g. $\lambda_2(1)$ is close to zero.

Surprisingly, we also obtained very good results when we tried to choose simply $q_2 = \hat{w}$. Then we have the shifted Broyden update (3.4) with adding term $-v_2v_2^T/(\bar{a}\bar{\delta})$; matrix $v_2v_2^T$ seems to have similar properties as $(as - bHy)(as - bHy)^T$ in case of the Broyden class, see [9]. Note that

$$v_2 = as - bHy - \zeta\left(\hat{a}s - \hat{b}Hy + \bar{a}Bs - \bar{b}y\right).$$

A method nearest to the Broyden class

The Broyden update, see [9], can be written in the form, similar to (3.4)

$$\frac{1}{\gamma}H_+^B = H + \frac{\varrho}{\gamma}\frac{ss^T}{b} - \frac{Hyy^TH}{a} + \frac{\eta}{a}\left(\frac{a}{b}s - Hy\right)\left(\frac{a}{b}s - Hy\right)^T. \tag{4.44}$$

Denoting $\bar{q}_2 = q_2/\sqrt{\bar{a}\bar{\delta}}$ and $M_3 = (1/\gamma)\left(H_+^B - (A_+ + \zeta_+ I)\right) + \bar{q}_2\bar{q}_2^T$, where A_+ is given by (4.24), we will seek to find $\bar{q}_2$ satisfying the following problem

$$\bar{q}_2 = \arg\min\{\|M_3 - qq^T\|_F^2 : \; q \in \mathcal{R}^N\}, \qquad \text{s.t.} \;\; q^Ty = 0. \tag{4.45}$$

It follows from (4.44) and (4.24) that

$$M_3 = \lambda_0 I + \frac{\varrho}{\gamma}\left(\frac{ss^T}{b} - \frac{\tilde{s}\tilde{s}^T}{\tilde{b}}\right) + \frac{Ayy^TA}{\bar{a}} - \frac{Hyy^TH}{a} + \frac{\eta}{a}\left(\frac{a}{b}s - Hy\right)\left(\frac{a}{b}s - Hy\right)^T + \frac{v_2v_2^T}{\bar{a}\bar{\delta}},$$

where $\lambda_0 = \zeta - (\varrho/\gamma)\sigma$. Using identities

$$\frac{ss^T}{b} - \frac{\tilde{s}\tilde{s}^T}{\tilde{b}} = \sigma\left(\frac{yy^T}{\hat{a}} - \hat{a}\frac{r_1r_1^T}{b\tilde{b}}\right), \qquad \frac{Ayy^TA}{\bar{a}} - \frac{Hyy^TH}{a} = \zeta\left(\hat{a}\frac{r_2r_2^T}{a\bar{a}} - \frac{yy^T}{\hat{a}}\right),$$

where $r_1 = s - (b/\hat{a})y$, $r_2 = Hy - (a/\hat{a})y = Ay - (\bar{a}/\hat{a})y$, we have

$$M_3 = \lambda_0\left(I - \frac{yy^T}{\hat{a}}\right) - \mu\frac{\varrho}{\gamma}\frac{r_1r_1^T}{\tilde{b}} + \zeta\hat{a}\frac{r_2r_2^T}{a\bar{a}} + \frac{\eta}{a}\left(\frac{a}{b}s - Hy\right)\left(\frac{a}{b}s - Hy\right)^T + \frac{v_2v_2^T}{\bar{a}\bar{\delta}}. \tag{4.46}$$

To solve problem (4.45), we utilize Lemma 4.3. First we readily deduce from (4.46) that every eigenvector of M_3 is a linear combination of vectors s, Hy, Bs and y. Since $M_3 y = 0$ by (4.46), y is the eigenvector corresponding to zero eigenvalue; thus any eigenvector corresponding to nonzero eigenvalue is perpendicular to y and therefore belongs to

$$\mathcal{P} = \Big\{ r : \; r \in \operatorname{span}\{s, Hy, Bs, y\}, \; r^T y = 0 \Big\} = \operatorname{span}\{r_1, r_2, r_3\},$$

where $r_3 = Bs - (\hat{b}/\hat{a})y$. Let Z be a matrix with i columns, $1 \le i \le 3$, creating an orthonormal basis in $\mathcal{P}$ (we still suppose $\bar{\delta} \neq 0$, which contradicts $r_1 = r_2 = 0$). Then $Z^T Z = I$ and Lemma 4.3 gives $M_3 \bar{q}_2 = |\bar{q}_2|^2 q_2$ and $\bar{q}_2 = Zh$ for some $h \in \mathcal{R}^i$, which yields

$$Z^T M_3 Z h = |h|^2 h. \tag{4.47}$$

Since $|\bar{q}_2| = |h|$, we will calculate the eigenvector h of $Z^T M_3 Z$, which corresponds to the largest eigenvalue of this matrix.

To construct a type 2 update, we first calculate vectors r_1, r_2 and r_3, orthonormalize them, create symmetric matrix $Z^T M_3 Z$ and calculate its eigenvalues and eigenvectors. Denoting λ_j, $j = 1, \ldots, i$, eigenvalues of $Z^T M_3 Z$ arranged in descending order and h_1 the eigenvector corresponding to λ_1, we calculate

$$q_2 = \sqrt{\bar{a}\bar{\delta}}\, \bar{q}_2 = \sqrt{\lambda_1 \bar{a}\bar{\delta}}\, Z h_1 / |h_1| \tag{4.48}$$

and then p_2 and p_1, using (4.26). In this connection, we have good experience with the Jacobi iteration method of finding eigenvalues and eigenvectors, which can also be utilized in the orthogonalization process to attain a high precision of results (if the columns of Q are eigenvectors of matrix $R^T R$, where $R = (r_1, r_2, r_3)$, then the columns of RQ create an orthogonal system and have norms equal to the square root of eigenvalues of $R^T R$). Note that the computation time required by the Jacobi method can be neglected for large N.

Since $\|Z^T M_3 Z - h_1 h_1^T\|_F^2 = \|Z^T M_3 Z\|_F^2 - |h_1|^4 = \sum_{j=1}^{i} \lambda_j^2 - \lambda_1^2 = \sum_{j=2}^{i} \lambda_j^2$ by (4.47) as in (4.36), we should choose parameters of the method in such a way to make $\sum_{j=2}^{i} \lambda_j^2$ as small as possible.

4.3 Global convergence

In this section we utilize the results obtained in Section 3.4. To establish global convergence, we can directly use Theorem 3.4. If condition (3.16) is not satisfied for the chosen constant C, we use some other update which fulfils the global convergency conditions (see below).

Note that in case $\bar{\delta} = 0$, when the particular methods described in the previous section cannot be used, condition (3.16) is also satisfied under the assumptions of Theorem 3.4 and $\gamma_k \le 1$, $k \ge 1$. This can be seen, observing that we use update $(1/\gamma)A_+ = A + (\varrho/\gamma)\tilde{s}\tilde{s}^T/\tilde{b} - Ayy^T A/\bar{a}$ for $\bar{a} \neq 0$ and $(1/\gamma)A_+ = A + (\varrho/\gamma)\tilde{s}\tilde{s}^T/\tilde{b} - ABss^T BA/\bar{c}$ for $\bar{c} \neq 0$ (we recall that $\bar{a} + \bar{c} > 0$ by (4.20)) and that $|\tilde{s}|^2/\tilde{b} \le 1/[(1-\overline{\mu})\underline{G}]$ (see the proof of Theorem 3.5).

We will show that the situation can be even better than in methods described in Section 3. We denote $\hat{w}$ the value $\sqrt{\eta\bar{\delta}}\left((\bar{a}/\tilde{b})\tilde{s} - Ay\right)$ as in Section 4.2.

Lemma 4.5. *Let $q_2 = \alpha\hat{w}+\beta v_2$ with $\alpha^2+\beta^2 \le 1$. Then the trace of matrix A_+ obtained by using update (4.24) cannot be greater than the trace of A_+ obtained by using update (3.4).*

Proof. One has

$$\begin{aligned}|q_2|^2 &= \alpha^2|\hat{w}|^2 + 2\alpha\beta v_2^T\hat{w} + \beta^2|v_2|^2 \le (1-\beta^2)|\hat{w}|^2 + 2\alpha\beta v_2^T\hat{w} + (1-\alpha^2)|v_2|^2 \\ &= |\hat{w}|^2 + |v_2|^2 - |\beta\hat{w} - \alpha v_2|^2 \le |\hat{w}|^2 + |v_2|^2,\end{aligned}$$

thus

$$\left(|q_2|^2 - |v_2|^2\right)/\bar{\delta} \le |\hat{w}|^2/\bar{\delta} = \eta\left|(\bar{a}/\tilde{b})\tilde{s} - Ay\right|^2,$$

which gives the desired result. □

Theorem 4.2. *The following three methods described in Section 4.2 satisfy the assumptions of Lemma 4.5 with $\alpha^2+\beta^2 = 1$: the method with direction vector derived from the shifted Broyden class, the method nearest to the shifted Broyden class and the method with $q_2 = \hat{w}$.*

Proof. In case of the first method, it follows from (4.33)-(4.34) that

$$\alpha = \hat{w}^T Bs/\sqrt{\bar{\delta}^2 + (\hat{w}^T Bs)^2}, \qquad \beta = \bar{\delta}/\sqrt{\bar{\delta}^2 + (\hat{w}^T Bs)^2}.$$

As regards the second method, q_2 is the eigenvector of $M_2 = \hat{w}\hat{w}^T + v_2 v_2^T$ corresponding to the nonzero eigenvalue λ_1, $|q_2|^2 = \lambda_1$ by Lemma 4.3. Then

$$q_2 = \frac{\hat{w}^T q_2}{\lambda_1}\hat{w} + \frac{v_2^T q_2}{\lambda_1}v_2, \qquad \alpha^2+\beta^2 = \frac{(\hat{w}^T q_2)^2 + (v_2^T q_2)^2}{\lambda_1^2} = \frac{q_2^T q_2}{\lambda_1} = 1.$$

Proof for the third method is obvious. □

In Section 3.4 we described the hybrid globally convergent shifted VM method, from which also limited memory globally convergent methods can be derived owing to Theorem 4.2.

4.4 Computational experiments

Similarly as in Section 3.6, the limited-memory VM methods were tested, using the collection of relatively difficult problems with optional dimension chosen from [10], [12] and [15]. We have used $\varrho = \gamma = 1$, $m = 10$ for $N = 50$ or $m = 20$ otherwise, the final precision $|g(x^\star)| \le 10^{-6}$ with η of the corresponding shifted Broyden class equal to unit and the choice (3.11) of the shift parameter μ in all iterations (except for methods SBC and NBC, see below). For starting iterates we use the shifted BFGS method as in Section 3.6. Results of our experiments are given in three tables, for $N = 50$, 200 and 1000, where NIT is the total number of iterations (over all problems) and NFV the total number of function evaluations. 'Fail' denotes the number of problems which were not solved successfully (usually NFV reached its limit).

The first six rows of tables give results for various methods described in Section 4: T1 – type 1 method (4.17), SBC – the simple method (4.29)-(4.30), SNSBC – the simplified variant of NSBC with $q_2 = \hat{w}$, NSBC – the method nearest to the shifted

Broyden class, DVSBC – the method (4.33)-(4.34) with direction vector derived from the shifted Broyden class and NBC – the method nearest to the Broyden class with μ and η obtained by quadratic interpolation.

For comparison, the last three rows contain results for the following limited-memory VM methods with 10 stored vectors for $N = 50$ or 20 vectors otherwise: RH – the reduced-Hessian method described in [5], BNS – the method after [1] and STRANG – the method based on the Strang formula, see [13]. Note that methods BNS and STRANG store pairs of vectors, here 5 pairs for $N = 50$ and 10 pairs otherwise.

Method	NIT	NFV	Fail
T1	19743	20798	-
SBC	23980	24971	2
SNSBC	18618	19546	-
NSBC	16486	17522	-
DVSBC	15575	16497	-
NBC	16725	17929	-
RH	22378	26801	-
BNS	25038	27792	-
STRANG	23754	26273	-

Table 1 ($N = 50$, 89 problems)

Method	NIT	NFV	Fail
T1	91722	96736	1
SBC	76960	79074	1
SNSBC	68289	71921	-
NSBC	76205	79371	-
DVSBC	74779	78738	-
NBC	64288	67877	-
RH	82267	93477	1
BNS	86690	97598	1
STRANG	86062	90957	1

Table 2 ($N = 200$, 88 problems)

Method	NIT	NFV	Fail
T1	23190	23627	-
SBC	22124	22321	-
SNSBC	17792	18009	-
NSBC	20236	20652	-
DVSBC	18364	18580	-
NBC	19298	20060	-
RH	21712	33314	-
BNS	18564	24747	1
STRANG	20195	21231	-

Table 3 ($N = 1000$, 22 problems)

Bibliography

[1] R.H. Byrd, J. Nocedal, R.B. Schnabel: *Representation of quasi-Newton matrices and their use in limited memory methods*, Math. Programming 63 (1994) 129-156.

[2] J.E.Dennis, J.J.Moré: *A characterization of superlinear convergence and its application to quasi-Newton methods*, Math. Comput. 28 (1974) 549-560.

[3] R. Fletcher: *Practical Methods of Optimization*, John Wiley & Sons, Chichester, 1987.

[4] P.E. Gill, M.W. Leonard: *Reduced-Hessian quasi-Newton methods for unconstrained optimization*, SIAM J. on Optimization, 12 (2001), 209-237.

[5] P.E. Gill, M.W. Leonard: *Limited-memory reduced-Hessian methods for large-scale unconstrained optimization*, Report NA 97-1 (revised), Dept of Mathematics, Santa Clara University, Santa Clara, 2002.

[6] G.H. Golub, C.Van Loan: *Matrix Computations* (Academic Press, NY, 1981).

[7] M.W. Leonard: *Reduced Hessian quasi-Newton methods for optimization*, PhD thesis, Dept of Mathematics, University of California, San Diego, 1995.

[8] L. Lukšan: *Computational experience with known variable metric updates*, J. Optim. Theory Appl. 83 (1994) 27-47.

[9] L. Lukšan, E. Spedicato: *Variable metric methods for unconstrained optimization and nonlinear least squares*, J. Comput. Appl. Math. 124 (2000) 61-95.

[10] L. Lukšan, J. Vlček: *Sparse and partially separable test problems for unconstrained and equality constrained optimization*, Report V-767, Prague, ICS AS CR, 1998.

[11] L. Lukšan, J. Vlček: *Metoda redukovaných Hessiánů pro nehladkou nepodmíněnou minimalizaci*, Programy a algoritmy num. mat. 10, Lázně Libverda, 2000.

[12] J.J. Moré, B.S. Garbow, K.E. Hillström: *Testing Unconstrained Optimization Software*, ACM Trans. Math. Software 7 (1981) 17-41.

[13] J. Nocedal: *Updating quasi-Newton matrices with limited storage*, Math. Comp. 35 (1980) 773-782.

[14] J. Nocedal, Y. Yuan: *Analysis of a Self-Scaling Quasi-Newton Method*, Math. Programming 61 (1993) 19-37.

[15] A. Roose, V. Kulla, M. Lomp, T. Meressco: *Test examples of systems of nonlinear equations*, Estonian Software and Computer Service Company, Tallin, 1990.

[16] D.F. Shanno, K.J. Phua: *Matrix conditioning and nonlinear optimization*, Math. Programming 14 (1978) 144-160.

[17] D. Siegel: *Implementing and modifying Broyden class updates for large scale optimization*, Report DAMTP NA12, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, 1992.