Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Computational and Applied Mathematics 186 (2006) 365–390

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

www.elsevier.com/locate/cam

Shifted limited-memory variable metric methods for large-scale unconstrained optimization☆

Jan Vlček[a], Ladislav Lukšan[a, b, *]

[a] *Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic*
[b] *Technical University of Liberec, Hálkova 6, 461 17 Liberec, Czech Republic*

Received 15 June 2004; received in revised form 26 January 2005

Abstract

A new family of numerically efficient full-memory variable metric or quasi-Newton methods for unconstrained minimization is given, which give simple possibility to derive related limited-memory methods. Global convergence of the methods can be established for convex sufficiently smooth functions. Numerical experience by comparison with standard methods is encouraging.

Keywords: Unconstrained minimization; Variable metric methods; Limited-memory methods; Global convergence; Numerical results

1. Introduction

Basic optimization methods can be realized in various ways which differ in direction determination and step-size selection. For unconstrained minimization of medium-size problems, variable metric (VM)

☆ This work was supported by the Grant Agency of the Czech Academy of Sciences, project code IAA1030405, and partly supported by the Institutional Research Plan AV0Z10300504 "Computer Science for the Information Society: Models, Algorithms, Applications".

* Corresponding author. Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic. Tel.: +420 266 053 260; fax: +420 286 585 789.

E-mail address: luksan@cs.cas.cz (L. Lukšan).

0377-0427/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.cam.2005.02.010

methods (see [5,10]) are most popular because of their stability and efficiency. Starting with an initial point $x_1 \in \mathscr{R}^N$, they generate a sequence $x_k \in \mathscr{R}^N$, by the process $x_{k+1} = x_k + t_k d_k$, $k \geqslant 1$, where

$$d_k = -H_k g_k \tag{1}$$

is a direction vector, t_k is a step-size and H_k is a symmetric positive definite matrix.

We will assume that the problem function $f : \mathscr{R}^N \to \mathscr{R}$ is differentiable and denote $f_k = f(x_k)$, $g_k = \nabla f(x_k)$, $s_k = x_{k+1} - x_k = t_k d_k$ and $y_k = g_{k+1} - g_k$, $k \geqslant 1$. We will investigate line-search methods with the step-size $t_k > 0$ chosen in such a way that

$$f_{k+1} - f_k \leqslant \varepsilon_1 t_k g_k^{\mathrm{T}} d_k, \quad g_{k+1}^{\mathrm{T}} d_k \geqslant \varepsilon_2 g_k^{\mathrm{T}} d_k, \tag{2}$$

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

Important property of the line-search method is the global convergence defined by relation

$$\liminf_{k \to \infty} |g_k| = 0. \tag{3}$$

The following theorem, see [5,10], characterizes the global convergence of the line-search method.

Theorem 1.1. *Let the objective function* $f : \mathscr{R}^N \to \mathscr{R}$ *be bounded from below and have bounded second-order derivatives. Consider the line-search method satisfying* (2). *If*

$$\sum_{k=1}^{\infty} \cos^2 \theta_k \triangleq \sum_{k=1}^{\infty} \frac{(g_k^{\mathrm{T}} d_k)^2}{g_k^{\mathrm{T}} g_k d_k^{\mathrm{T}} d_k} = \infty \tag{4}$$

and $g_k^{\mathrm{T}} d_k < 0$, $k \geqslant 1$, *then* (3) *holds.*

Our work was motivated by an effort to develop efficient methods for large-scale unconstrained optimization. Standard VM methods use dense matrices which are updated in every iteration. This is unsuitable and often impossible, when the number of variables is large. Therefore, matrix-free methods have been developed, which eliminate this insufficiency. Conjugate gradient methods form a simplest class of such methods, but their rate of convergence is usually rather slow in comparison with variable metric methods and also the final accuracy obtained is not always quite sufficient. Therefore, new principles based on variable metric updates were sought. The limited-memory BFGS method [13] was the first one which uses variable metric updates in the vector form (the so called Strang formula). Later a compact form utilizing small-size matrices was proposed in [3]. These methods use $m \ll N$ pairs of vectors s_i, y_i, $k - m \leqslant i \leqslant k - 1$, in kth iteration and construct matrix H_k by m variable matric updates from the scaled unit matrix. Therefore, information obtained in iterations with indices lower than $k - m$ is completely lost. Limited-memory methods of this type were later modified and improved, e.g., in papers [8] and [1].

Recently a different principle based on reduced Hessian matrices was introduced in [7]. In this case, only m vectors s_i, $k - m \leqslant i \leqslant k - 1$, are saved and the approximation of the inverse Hessian matrix has the form $H_k = Z_k (Z_k^{\mathrm{T}} B_k Z_k)^{-1} Z_k^{\mathrm{T}}$, where Z_k is a matrix whose orthonormal columns form a basis in the subspace spanned by vectors s_i, $k - m \leqslant i \leqslant k - 1$, and where $Z_k^{\mathrm{T}} B_k Z_k$ is an approximation of the small-size reduced Hessian matrix, which is updated by variable metric updates. Since the number of columns of Z_k is limited, the oldest column is usually discarded in kth iteration. Thus a part of information is again lost. Moreover, matrix H_k, which can be written in the form $H_k = U_k U_k^{\mathrm{T}}$, where U_k is a rectangular matrix, is singular. Thus the case when d_k is small or almost perpendicular to g_k can occur after discarding columns from Z_k. For this reason, we decided to use matrix of the form $H_k = \zeta_k I + U_k U_k^{\mathrm{T}}$, where $\zeta_k > 0$ and U_k is a rectangular matrix with m columns, which is updated in every iteration in such a way that no information is discarded. The choice of parameter ζ_k is of course crucial (see Section 2.2). We call these methods the shifted limited-memory VM methods.

Since these methods need a suitable starting matrix U_m, we have developed full-memory shifted VM methods as alternative to the well-known standard Broyden class of VM methods, see e.g. [5], which increase number of columns of U_k by 1 in every update. In Section 2 we describe particular methods of this type and give a numerical comparison with the standard VM methods.

Section 3 is devoted to the shifted limited-memory VM methods. We give description of particular methods, including variationally-derived methods and numerical results, which confirm their efficiency and stability. In Section 4 we establish global convergence of our methods for f uniformly convex and describe a simple way allowing to develop globally convergent methods in the nonconvex case.

2. Shifted variable metric methods

Variable metric methods, see [5,10], use symmetric positive definite matrices H_k or $B_k = H_k^{-1}$, $k \geqslant 1$; usually $H_1 = I$ and H_{k+1} is obtained from H_k by a rank-two VM update to satisfy the quasi-Newton condition $H_{k+1} y_k = s_k$.

In shifted VM methods, matrices H_k have the form

$$H_k = \zeta_k I + A_k, \tag{5}$$

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

$$A_{k+1} y_k = \varrho_k \tilde{s}_k, \quad \tilde{s}_k = s_k - \zeta_{k+1} y_k, \tag{6}$$

where parameter $\varrho_k > 0$ represents analogy of nonquadratic correction, see [2,10], but since it is used with matrices A_k instead of H_k, its influence and methods of calculation are quite different. Note that neither using of this correction parameter in a standard way (with matrices H_k), nor standard scaling, see [10,14], improved our results substantially and we do not use them in this paper. If $\varrho_k = 1$, relations (5), (6) obviously imply that matrix H_{k+1} satisfies the quasi-Newton condition $H_{k+1} y_k = s_k$. Note that we use non-unit values of ϱ_k in our numerical experiments only for variationally-derived limited-memory methods (see Section 3.2).

To simplify the notation we often omit index k and replace index $k+1$ by symbol $+$. In the subsequent analysis we use the following notation:

$$a = y^{\mathrm{T}} H y, \quad \bar{a} = y^{\mathrm{T}} A y, \quad \hat{a} = y^{\mathrm{T}} y, \quad b = s^{\mathrm{T}} y, \quad \bar{b} = s^{\mathrm{T}} B A y, \quad \tilde{b} = \tilde{s}^{\mathrm{T}} y, \quad \bar{c} = s^{\mathrm{T}} B A B s.$$

In this section we concentrate on the shifted analogy of the Broyden class, see [5,10]. Using the same argumentation as in standard VM methods, we consider the shifted VM update for $\tilde{b} > 0$ (which implies $\tilde{s} \neq 0$, $y \neq 0$) in the form

$$A_+ = A + \varrho \frac{\tilde{s}\tilde{s}^{\mathrm{T}}}{\tilde{b}} - \frac{Ayy^{\mathrm{T}}A}{\bar{a}} + \frac{\eta}{\bar{a}} \left(\frac{\bar{a}}{\tilde{b}} \tilde{s} - Ay \right) \left(\frac{\bar{a}}{\tilde{b}} \tilde{s} - Ay \right)^{\mathrm{T}}, \tag{7}$$

(if $\bar{a} = 0$, i.e. $Ay = 0$ by $\bar{a} = |A^{1/2}y|^2$, we simply omit the last two terms, because their limit value is zero for $Ay = \lim_{\xi \to 0} \xi q$, $\bar{a} = \lim_{\xi \to 0} \xi q^{\mathrm{T}} y$, $q^{\mathrm{T}} y \neq 0$; in this case the update is independent of η), 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

$$A_+^{s\mathrm{DFP}} = A + \varrho \frac{\tilde{s}\tilde{s}^{\mathrm{T}}}{\tilde{b}} - \frac{Ayy^{\mathrm{T}}A}{\bar{a}}, \quad A_+^{s\mathrm{BFGS}} = A + \left(\varrho + \frac{\bar{a}}{\tilde{b}} \right) \frac{\tilde{s}\tilde{s}^{\mathrm{T}}}{\tilde{b}} - \frac{\tilde{s}y^{\mathrm{T}}A + Ay\tilde{s}^{\mathrm{T}}}{\tilde{b}}. \tag{8}$$

2.1. Basic properties

Theorem 2.1. *Let A be positive semidefinite, $\eta \geqslant 0$ and $\zeta_+ \hat{a} < b$. Then matrix A_+ given by* (7) *is positive semidefinite.*

Proof. Since $\tilde{b} = \tilde{s}^{\mathrm{T}} y = b - \zeta_+ \hat{a} > 0$ by (6), the positive semidefiniteness of matrix A_+ follows from (7) for $\bar{a} = 0$, otherwise from the quasi-product form of (7)

$$A_+ = \left(I - \left(\frac{\sqrt{\eta}}{\tilde{b}} \tilde{s} + \frac{1 - \sqrt{\eta}}{\bar{a}} Ay \right) y^{\mathrm{T}} \right) A \left(I - y \left(\frac{\sqrt{\eta}}{\tilde{b}} \tilde{s} + \frac{1 - \sqrt{\eta}}{\bar{a}} Ay \right)^{\mathrm{T}} \right) + \varrho \frac{\tilde{s}\tilde{s}^{\mathrm{T}}}{\tilde{b}}, \tag{9}$$

which can be readily verified, using straightforward arrangements and comparing corresponding terms. □

Note that for $\eta = 0$ we can write matrix A_+ in the product form

$$A_+^{s\mathrm{DFP}} = \left(I - \left(\pm\sqrt{\varrho \bar{a}/\tilde{b}}\tilde{s} + Ay \right) \frac{y^{\mathrm{T}}}{\bar{a}} \right) A \left(I - \frac{y}{\bar{a}} \left(\pm\sqrt{\varrho \bar{a}/\tilde{b}}\tilde{s} + Ay \right)^{\mathrm{T}} \right). \tag{10}$$

From now on we will suppose that $\eta \geqslant 0$ and $\tilde{b} > 0$. In view of Theorem 2.1, the shift parameter ζ_+ should satisfy inequality $0 < \zeta_+ < b/\hat{a}$. Therefore, it is advantageous to introduce relative shift parameter $\mu = \zeta_+ \hat{a}/b \in (0, 1)$ and by (6) we can write

$$\zeta_+ = \mu b/\hat{a}, \quad \tilde{b} = \tilde{s}^{\mathrm{T}} y = b - \zeta_+ \hat{a} = b(1 - \mu). \tag{11}$$

2.2. Determination of the shift parameter

Determination of the shift parameter μ (or ζ_+) is a crucial part of the shifted VM method because the choice of ζ_+ influences the lowest eigenvalue of matrix H_+. Therefore μ should not be close to zero when matrix A is not sufficiently positive definite. On the other hand, $\|A_+\|$ can increase explosively when μ tends to unit (see below).

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

$$0.20 \leqslant \mu \leqslant 0.25, \tag{12}$$

(e.g., the choice $\mu = 0.22$) appear to be suitable in this case. If $\mu \geqslant 1/2$, then the convergence is usually lost, see Section 2.4 (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 $A = UU^{\mathrm{T}}$, where U is a rectangular matrix, gives a useful formula for determination of parameter μ.

Lemma 2.1. *Denoting $\nu = \mu/(1-\mu)$, $\phi = \nu\sqrt{1 - b^2/(\hat{a}|s|^2)}$, $V = I - sy^{\mathrm{T}}/b$ and $\tilde{V} = I - \tilde{s}y^{\mathrm{T}}/\tilde{b}$, there holds $\|\tilde{V} - V\|/\|V\| = \phi$. Moreover, let vector $u \in \mathscr{R}^N$, $y^{\mathrm{T}}u \neq 0$, be scaled to satisfy $y^{\mathrm{T}}u = b$. Then*

$$\phi - (1+\phi)|u-s|/|u| \leqslant |\tilde{V}u|/|u| \leqslant \phi + (1+\phi)|u-s|/|u|. \tag{13}$$

Proof. One has $\tilde{s} = s - \mu(b/\hat{a})y$ and $\tilde{b} = (1-\mu)b$ by (6) and (11) and thus

$$\tilde{V} - V = \frac{(1-\mu)s - s + \mu(b/\hat{a})y}{(1-\mu)b}\, y^{\mathrm{T}} = \frac{-\mu[s - (b/\hat{a})y]}{(1-\mu)b}\, y^{\mathrm{T}} = -\frac{\nu}{b}\left(s - \frac{b}{\hat{a}}\, y\right) y^{\mathrm{T}}.$$

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

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

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

Let $y^{\mathrm{T}}u = b$. By (6) and (11) 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:

$$|\tilde{V}u| \leqslant \phi|s| + |u-s| \leqslant \phi(|u| + |u-s|) + |u-s| = \phi|u| + (1+\phi)|u-s|,$$

$$|\tilde{V}u| \geqslant \phi|s| - |u-s| \geqslant \phi(|u| - |u-s|) - |u-s| = \phi|u| - (1+\phi)|u-s|. \qquad \square$$

Now we turn back to the shift parameter determination. Value $\|\tilde{V} - V\|/\|V\|$, equal to ϕ by Lemma 2.1, represents a relative deviation of $\tilde{V}$ from V. The shifted BFGS update $A_+ = \tilde{V}UU^{\mathrm{T}}\tilde{V}^{\mathrm{T}} + \varrho\tilde{s}\tilde{s}^{\mathrm{T}}/\tilde{b}$, see (9), multiplies columns of U by $\tilde{V}$. In the BFGS update, see [10], which can be written in the form $H_+ = VHV^{\mathrm{T}} + ss^{\mathrm{T}}/b$, multiplication by V instead of $\tilde{V}$ is performed. Thus in case $A \approx H$ and if $\|A\|$ is great compared to $\|\varrho\tilde{s}\tilde{s}^{\mathrm{T}}/\tilde{b} - ss^{\mathrm{T}}/b\|$, 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, $\bar{u} = (b/u^{\mathrm{T}}y)u$ for $u^{\mathrm{T}}y \neq 0$, we have $s \approx \xi u$ for some $\xi \in \mathscr{R}$ by (1), thus $s \approx \bar{u}$ and by (13) we get

$|\tilde{V}u|/|u| = |\tilde{V}\bar{u}|/|\bar{u}| \approx \phi$. Therefore for $\phi > 1$ 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 $\phi = 1$, we obtain $\mu_1 = 1/(1 + \sqrt{1 - b^2/(\hat{a}|s|^2)})$. This value can serve as a reasonable maximum of μ and should be multiplied by coefficient $\varepsilon > 0$ with the properties

- if $U^{\mathrm{T}}y = 0$ then $\varepsilon = 1$ because $\tilde{V}U = U$ and it is not necessary to decrease μ,
- if $\bar{a} = |U^{\mathrm{T}}y|^2 > 0$ 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 ε satisfies conditions for global convergence of the shifted BFGS method (see Theorem 4.2). Multiplying μ_1 by ε, we obtain finally

$$\mu = \sqrt{1 - \bar{a}/a} \bigg/ \left(1 + \sqrt{1 - b^2/(\hat{a}|s|^2)}\right). \tag{14}$$

In the first iteration, this value of μ has the following interesting property.

Theorem 2.2. *Let* $A = 0$. *Then matrix* $H_+ = \zeta_+ I + A_+$ *with value* (14), *where* A_+ *is given by* (7), *is optimally conditioned.*

Proof. If $A = 0$, thus $\bar{a} = 0$, formula (7) (where we omit the last two terms) gives $H_+ = \zeta_+ I + \varrho\tilde{s}\tilde{s}^{\mathrm{T}}/\tilde{b}$, which yields $H_+^{-1} = (1/\zeta_+)[I - \varrho\tilde{s}\tilde{s}^{\mathrm{T}}/(\zeta_+\tilde{b} + \varrho|\tilde{s}|^2)]$. Thus $\|H_+\| = \zeta_+ + \varrho|\tilde{s}|^2/\tilde{b}$, $\|H_+^{-1}\| = 1/\zeta_+$ and $\kappa_+ \triangleq \|H_+\|\|H_+^{-1}\| = 1 + \varrho|\tilde{s}|^2/(\zeta_+\tilde{b})$. By (6), (11) and denoting again $v = \mu/(1 - \mu)$, we obtain

$$\frac{\kappa_+ - 1}{\varrho} = \frac{|\tilde{s}|^2}{\mu(1-\mu)b^2/\hat{a}} = \frac{\hat{a}}{vb^2}\left|\frac{s - \mu(b/\hat{a})y}{1-\mu}\right|^2 = \frac{\hat{a}}{vb^2}\left|s(1+v) - v\frac{b}{\hat{a}}y\right|^2$$
$$= \frac{\hat{a}}{vb^2}\left(|s|^2(1+v)^2 - \frac{b^2}{\hat{a}}(v^2 + 2v)\right) = \frac{\hat{a}}{vb^2}|s|^2 + (v+2)\left(\frac{\hat{a}}{b^2}|s|^2 - 1\right),$$

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

$$(\hat{a}/b^2)|s|^2(1 - 1/v^2) = 1$$

with the positive root $v = 1/\sqrt{1 - b^2/(\hat{a}|s|^2)}$. By $\bar{a} = 0$, this corresponds to (14). □

Formula (14) gives good results with update (7) without any corrections, with the exception of the first 5 to 10 iterations, when it should be corrected, e.g., in the following way:

$$\mu = \min\left(\max\left(\sqrt{1 - \bar{a}/a} \bigg/ \left(1 + \sqrt{1 - b^2/(\hat{a}|s|^2)}\right), 0.2\right), 0.8\right), \tag{15}$$

because our reasoning leading to (14) was simplified and the shifted VM methods effectivity is very sensitive to the shift parameter determination in the first iterations.

2.3. The shifted DFP method

Starting with $A=0$, (8) gives $A_+^{s\mathrm{DFP}}=\varrho\tilde{s}\tilde{s}^{\mathrm{T}}/\tilde{b}$. The following theorem shows that this form of A_+, which needs no matrix storage, is typical for the shifted DFP method.

Theorem 2.3. *Let $A=uu^{\mathrm{T}}$, $u^{\mathrm{T}}y\neq 0$. Then $A_+^{s\mathrm{DFP}}=\varrho\tilde{s}\tilde{s}^{\mathrm{T}}/\tilde{b}$.*

Proof. Since $Ay=(u^{\mathrm{T}}y)u$, we obtain from (8)

$$A_+^{s\mathrm{DFP}}=uu^{\mathrm{T}}+\varrho\,\frac{\tilde{s}\tilde{s}^{\mathrm{T}}}{\tilde{b}}-(u^{\mathrm{T}}y)^2\,\frac{uu^{\mathrm{T}}}{(u^{\mathrm{T}}y)^2}=\varrho\,\frac{\tilde{s}\tilde{s}^{\mathrm{T}}}{\tilde{b}}.\qquad\square$$

This result can be generalized for rank-two matrix A.

Theorem 2.4. *Let $A=u_1u_1^{\mathrm{T}}+u_2u_2^{\mathrm{T}}$, $v_2=\bar{a}ABs-\bar{b}Ay$, $\bar{\delta}\triangleq\bar{a}\bar{c}-\bar{b}^2\neq 0$. Then*

$$A_+^{s\mathrm{DFP}}=\varrho\,\frac{\tilde{s}\tilde{s}^{\mathrm{T}}}{\tilde{b}}+\frac{v_2v_2^{\mathrm{T}}}{\bar{a}\bar{\delta}}.\tag{16}$$

Proof. It follows from the Schwarz inequality that $\bar{\delta}\geqslant 0$, thus $\bar{\delta}\neq 0$ implies $\bar{a}\neq 0$. Denoting $\alpha_i=u_i^{\mathrm{T}}y$, $\beta_i=u_i^{\mathrm{T}}Bs$, $i=1,2$, we obtain $Ay=\alpha_1u_1+\alpha_2u_2$, $\bar{a}=\alpha_1^2+\alpha_2^2$ and similar relations for ABs, $\bar{b}$ and $\bar{c}$. Therefore

$$\bar{\delta}=(\alpha_1^2+\alpha_2^2)(\beta_1^2+\beta_2^2)-(\alpha_1\beta_1+\alpha_2\beta_2)^2=(\alpha_2\beta_1-\alpha_1\beta_2)^2,$$

$$\begin{aligned}v_2&=(\alpha_1^2+\alpha_2^2)(\beta_1u_1+\beta_2u_2)-(\alpha_1\beta_1+\alpha_2\beta_2)(\alpha_1u_1+\alpha_2u_2)\\&=(\alpha_2\beta_1-\alpha_1\beta_2)(\alpha_2u_1-\alpha_1u_2).\end{aligned}$$

Since $A_+^{s\mathrm{DFP}}-\varrho\tilde{s}\tilde{s}^{\mathrm{T}}/\tilde{b}=A-(1/\bar{a})Ayy^{\mathrm{T}}A$ by (8), we have

$$\begin{aligned}\bar{a}(A_+^{s\mathrm{DFP}}-\varrho\tilde{s}\tilde{s}^{\mathrm{T}}/\tilde{b})&=(\alpha_1^2+\alpha_2^2)(u_1u_1^{\mathrm{T}}+u_2u_2^{\mathrm{T}})-(\alpha_1u_1+\alpha_2u_2)(\alpha_1u_1+\alpha_2u_2)^{\mathrm{T}}\\&=(\alpha_2u_1-\alpha_1u_2)(\alpha_2u_1-\alpha_1u_2)^{\mathrm{T}}=v_2v_2^{\mathrm{T}}/\bar{\delta}.\qquad\square\end{aligned}$$

The product form (10) shows that for $\bar{a}\neq 0$ the rank of the updated matrix cannot increase. Thus this method does not accumulate information from previous iterations sufficiently, which probably causes its less efficiency.

2.4. Computational experiments

The shifted VM methods were tested using a collection of 92 relatively difficult problems with optional dimension chosen from [12], which can be downloaded from the web page `http://www.cs.cas.cz/~luksan/test.html` as TEST28. The results of our experiments are given in two tables, where NIT is the total number of iterations (over all 92 problems), NFV the total number of function (or gradient) evaluations and 'Fail' denotes the number of problems which were not solved successfully (usually NFV reached its limit). We have used dimensions $N=50,\ 200$ and the final precision $\|g(x^{\star})\|_\infty\leqslant 10^{-6}$.

Table 1
$N = 50$

μ	NIT	NFV	Fail	Time
0.22	12222	13929	—	0.91
0.32	12617	15540	1	0.97
0.42	12874	18256	2	1.08
0.48	15994	28264	3	1.52
0.50	31118	65567	12	3.39
0.52	24947	102302	45	6.00

Table 2

Method	$N = 50$				$N = 200$			
	NIT	NFV	Fail	Time	NIT	NFV	Fail	Time
SBFGS	11449	12465	—	0.92	29864	34768	1	10.75
SDFP	46010	48579	9	3.30	81279	87624	19	27.38
SBC2	10997	12616	—	0.76	31651	38346	3	11.42
SHOS	13814	14716	—	0.92	36167	40660	3	12.41
BFGS	15170	16824	1	1.14	34725	38456	3	11.92
DFP/1	79873	84546	36	4.25	124040	136144	33	52.06
DFP/2	15560	36345	2	1.45	33524	76279	4	16.99
BC2	12566	14949	1	0.92	29072	34793	3	10.08
HOS	18529	19571	1	1.06	40453	42783	3	13.13

Table 1 demonstrates an influence of the constant parameter μ on the efficiency of the shifted BFGS method (the value 0.22 is in range (12)). We see that the convergence is lost when $\mu \geqslant 1/2$. In the next table we use choice (14) of the shift parameter μ with corrections (15) in the first six iterations.

The first five rows of Table 2 contain results for the following shifted VM methods: the shifted BFGS method (SBFGS, $\eta = 1$), the shifted DFP method (SDFP, $\eta = 0$) and method (7) with $\eta = 2$ (SBC2) and $\eta = b/(a + b)$ (SHOS, shifted analogy of Hoshino self-dual method, see [10]).

For comparison, the last five rows of the table contain results for various standard VM methods: the BFGS method with scaling in the first iteration (BFGS, see [15]), the DFP method without scaling (DFP/1), the DFP method without scaling with the strong Wolfe line-search conditions, where the second inequality in (2) is replaced by $|g_{k+1}^{\mathrm{T}} d_k| \leqslant \varepsilon_2 |g_k^{\mathrm{T}} d_k|$ with $\varepsilon_2 = 0.1$ (DFP/2), method from the Broyden class with $\eta = 2$ (BC2) and Hoshino self-dual method (HOS), both with scaling in the first iteration.

This table demonstrates the high efficiency of the shifted BFGS method. It is more efficient than the standard BFGS method with usual scaling strategies (other scaling strategies that can improve the efficiency of standard VM methods are introduced in [10]). Moreover, the modified shifted DFP method can give much more better results than the shifted DFP method and the shifted DFP method is much more efficient than the standard DFP method with usual scaling strategies and usual line-search methods.

3. Limited-memory methods

All methods investigated in this section belong to shifted VM methods. They satisfy (5)–(6) and (11) with (positive semidefinite) matrix $A_k = U_k U_k^{\mathrm{T}}$, where U_k, $k \geqslant 1$, is a rectangular matrix. Thus we store and update only matrix U_k. We again often omit index k and replace index $k+1$ by symbol $+$.

The shifted VM methods presented in Section 2, particularly in the quasi-product form (9), are ideal as starting methods. Setting $U_+ = (\sqrt{1/\tilde{b}}\tilde{s})$ in the first iteration, every update (9) modifies U and adds one column $\sqrt{1/\tilde{b}}\tilde{s}$ to U_+. Thus in this section we will assume that the starting iterations have been executed and that matrix U has $m \geqslant 1$ columns in all iterations.

We say that the method is of type i when the rank of matrix $U_+ - U$ is i, $i \geqslant 1$. The type 1 methods are simpler, but the type 2 methods appear to be more efficient in practice. The shifted DFP method (10) is an example of type 1 method. Better results were obtained with type 1 update formulas $U_+ = U + p(Bs + \vartheta y)^{\mathrm{T}} U = (I + \vartheta p y^{\mathrm{T}} + p s^{\mathrm{T}} B) U$ for suitable $p \in \mathscr{R}^N$ and $\vartheta \in \mathscr{R}$. To have more free parameters, we will investigate the following basic form of update:

$$U_+ = (I + p_1 y^{\mathrm{T}} + p_2 s^{\mathrm{T}} B) U, \quad p_1 \in \mathscr{R}^N, \ p_2 \in \mathscr{R}^N. \tag{17}$$

3.1. *Methods based on general expression of the basic update*

Many update formulas can be constructed by comparison of basic update (17) with the shifted Broyden class. To make this, it is useful to express update (17) in the form similar to (7). From (17) we have

$$A_+ = A + p_1 y^{\mathrm{T}} A + A y p_1^{\mathrm{T}} + p_2 s^{\mathrm{T}} B A + A B s p_2^{\mathrm{T}} + \bar{a} p_1 p_1^{\mathrm{T}} + \bar{b}(p_1 p_2^{\mathrm{T}} + p_2 p_1^{\mathrm{T}}) + \bar{c} p_2 p_2^{\mathrm{T}}. \tag{18}$$

Denoting $\tau_1 = 1 + p_1^{\mathrm{T}} y$, $\tau_2 = p_2^{\mathrm{T}} y$, the quasi-Newton condition (6) gives

$$(\bar{a}\tau_1 + \bar{b}\tau_2) p_1 + (\bar{b}\tau_1 + \bar{c}\tau_2) p_2 + \tau_1 A y + \tau_2 A B s = \varrho \tilde{s}, \tag{19}$$

$$\bar{a}\tau_1^2 + 2\bar{b}\tau_1\tau_2 + \bar{c}\tau_2^2 = \varrho \tilde{b}. \tag{20}$$

We will use the following notation (note that the Schwarz inequality implies $\bar{\delta} \geqslant 0$):

$$\bar{\delta} = \bar{a}\bar{c} - \bar{b}^2, \quad v_1 = \bar{c} A y - \bar{b} A B s, \quad v_2 = \bar{a} A B s - \bar{b} A y, \quad q_1 = \bar{\delta} p_1 + v_1, \quad q_2 = \bar{\delta} p_2 + v_2$$

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

$$q_i^{\mathrm{T}} y = \bar{\delta} \tau_i, \ i = 1, 2, \quad \bar{a}(v_1 v_1^{\mathrm{T}} + \bar{\delta} A B s s^{\mathrm{T}} B A) = \bar{c}(v_2 v_2^{\mathrm{T}} + \bar{\delta} A y y^{\mathrm{T}} A). \tag{21}$$

Lemma 3.1. *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. □

We still assume $\tilde{b} > 0$, thus at least one of values $\bar{a}, \bar{c}$ must be nonzero by (20) and $\bar{\delta} \geqslant 0$. First we will suppose that $\bar{a} \neq 0$ and that vectors p_1 and p_2 are chosen such that $\bar{a}\tau_1 + \bar{b}\tau_2 \neq 0$. Our approach is based on the following result.

Lemma 3.2. *Let* $\tilde{p} = \bar{a}p_1 + \bar{b}p_2$, $\omega_1 = \bar{a}\tau_1 + \bar{b}\tau_2$, $\bar{a}\omega_1 \neq 0$ *and let* (19) *hold. Then*

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

where

$$\hat{q}_2 = [q_2 - (q_2^{\mathrm{T}} y/\tilde{b})\tilde{s}]/(|\omega_1|\omega_2), \quad \omega_2 = 1/\sqrt{\varrho\bar{a}\tilde{b}}. \tag{22}$$

Proof. The first relation readily follows from (20), which is implied by (19). One has

$$\begin{aligned}\omega_1(\tilde{p} + Ay) &= (\bar{a}\tau_1 + \bar{b}\tau_2)(\bar{a}p_1 + \bar{b}p_2 + Ay) = \bar{a}(\varrho\tilde{s} - \tau_2(\bar{b}p_1 + \bar{c}p_2 + ABs)) \\ &\quad + \bar{b}\tau_2(\bar{a}p_1 + \bar{b}p_2 + Ay) = \varrho\bar{a}\tilde{s} - \tau_2\bar{\delta}p_2 - \tau_2 v_2 = \varrho\bar{a}\tilde{s} - \tau_2 q_2 \\ &= \varrho\bar{a}\tilde{s} - \tau_2(|\omega_1|\omega_2\hat{q}_2 + (\bar{\delta}\tau_2/\tilde{b})\tilde{s}) = |\omega_1|(|\omega_1|\tilde{s}/\tilde{b} - \tau_2\omega_2\hat{q}_2)\end{aligned} \tag{23}$$

by (19), (22) and $q_2^{\mathrm{T}} y = \bar{\delta}\tau_2$, thus $\tilde{p} + Ay = \pm(|\omega_1|\tilde{s}/\tilde{b} - \tau_2\omega_2\hat{q}_2)$ and

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

Before utilizing this lemma, we rewrite (18) in the following way:

$$\begin{aligned}\bar{a}(A_+ - A) &= \tilde{p}y^{\mathrm{T}}A + Ay\tilde{p}^{\mathrm{T}} + p_2 v_2^{\mathrm{T}} + v_2 p_2^{\mathrm{T}} + \tilde{p}\tilde{p}^{\mathrm{T}} + \bar{\delta}p_2 p_2^{\mathrm{T}} \\ &= p_2 v_2^{\mathrm{T}} + v_2 p_2^{\mathrm{T}} + (\tilde{p} + Ay)(\tilde{p} + Ay)^{\mathrm{T}} - Ayy^{\mathrm{T}}A + \bar{\delta}p_2 p_2^{\mathrm{T}}.\end{aligned} \tag{24}$$

Since $\bar{\delta}(p_2 v_2^{\mathrm{T}} + v_2 p_2^{\mathrm{T}}) + \bar{\delta}^2 p_2 p_2^{\mathrm{T}} = q_2 q_2^{\mathrm{T}} - v_2 v_2^{\mathrm{T}}$, we can use Lemma 3.2 to obtain $\bar{a}\bar{\delta}(A_+ - A) = (\varrho\bar{a}\bar{\delta}/\tilde{b})\tilde{s}\tilde{s}^{\mathrm{T}} - \bar{\delta}Ayy^{\mathrm{T}}A + \hat{q}_2\hat{q}_2^{\mathrm{T}} - v_2 v_2^{\mathrm{T}}$. Since $\hat{q}_2 = q_2$ 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 (20), and the update formula can be written in the form

$$A_+ = A + \varrho\frac{\tilde{s}\tilde{s}^{\mathrm{T}}}{\tilde{b}} - \frac{Ayy^{\mathrm{T}}A}{\bar{a}} + \frac{q_2 q_2^{\mathrm{T}} - v_2 v_2^{\mathrm{T}}}{\bar{a}\bar{\delta}}, \quad q_2^{\mathrm{T}} y = 0 \tag{25}$$

for $\bar{\delta} \neq 0$. If $\bar{\delta} = 0$, one has $v_2 = q_2 = \hat{q}_2 = 0$ by Lemma 3.1, thus $\tilde{p} + Ay = \omega_1\tilde{s}/\tilde{b}$ by (23) and (24) gives $\bar{a}(A_+ - A) = (\omega_1^2/\tilde{b}^2)\tilde{s}\tilde{s}^{\mathrm{T}} - Ayy^{\mathrm{T}}A$, therefore we get $A_+ = A + \varrho\tilde{s}\tilde{s}^{\mathrm{T}}/\tilde{b} - Ayy^{\mathrm{T}}A/\bar{a}$ (which is the shifted DFP update (8)) for any choice of p_2.

Proceeding similarly for $\bar{c} \neq 0$ (e.g., when $\bar{a} = 0$), we derive the following formula:

$$A_+ = A + \varrho \frac{\tilde{s}\tilde{s}^{\mathrm{T}}}{\tilde{b}} - \frac{ABss^{\mathrm{T}}BA}{\bar{c}} + \frac{q_1 q_1^{\mathrm{T}} - v_1 v_1^{\mathrm{T}}}{\bar{c}\bar{\delta}}, \quad q_1^{\mathrm{T}} y = 0 \tag{26}$$

for $\bar{\delta} \neq 0$ and $A_+ = A + \varrho\tilde{s}\tilde{s}^{\mathrm{T}}/\tilde{b} - ABss^{\mathrm{T}}BA/\bar{c}$ for $\bar{\delta} = 0$ (and any p_1); this update satisfies the shifted quasi-Newton condition by Lemma 3.1. By (21), update (26) can be for $\bar{a}\bar{c} \neq 0$ written (note that q_1 satisfying $q_1^{\mathrm{T}} y = 0$ cannot be proportional to q_2, $q_2^{\mathrm{T}} y = 0$, for $\bar{\delta} \neq 0$ by (21), since τ_1, τ_2 cannot be equal to zero simultaneously by (20))

$$A_+ = A + \varrho \frac{\tilde{s}\tilde{s}^{\mathrm{T}}}{\tilde{b}} - \frac{Ayy^{\mathrm{T}}A}{\bar{a}} + \frac{q_1 q_1^{\mathrm{T}}}{\bar{c}\bar{\delta}} - \frac{v_2 v_2^{\mathrm{T}}}{\bar{a}\bar{\delta}}, \quad q_1^{\mathrm{T}} y = 0. \tag{27}$$

Update formulas (25), (27) can be significantly simplified in case $m \leqslant 2$, using Theorem 2.4. Combining (8) and (16) with (25) and (27), we obtain the general form of type 1 or type 2 update for limited memory methods with $m \leqslant 2$ and $\bar{a}\bar{c} \neq 0$

$$A_+ = \varrho \frac{\tilde{s}\tilde{s}^{\mathrm{T}}}{\tilde{b}} + \frac{q_2 q_2^{\mathrm{T}}}{\bar{a}\bar{\delta}} \quad \text{or} \quad A_+ = \varrho \frac{\tilde{s}\tilde{s}^{\mathrm{T}}}{\tilde{b}} + \frac{q_1 q_1^{\mathrm{T}}}{\bar{c}\bar{\delta}}. \tag{28}$$

For example, the choice $q_2 = 0$ or $q_2 = v_2$ in the first formula gives the shifted DFP update for $m = 1$ or $m = 2$. This interesting formulas need not store any VM matrix, similarly as conjugate gradient methods, but can be more efficient.

To construct limited-memory update, we can proceed in the following way. If $\bar{\delta} \neq 0$ (thus also $\bar{a}\bar{c} \neq 0$ by $\bar{\delta} \geqslant 0$) we choose vector parameter q_2 satisfying $q_2^{\mathrm{T}} y = 0$, i.e. $\tau_2 = 0$. Then $\tau_1 = \pm\sqrt{\varrho\tilde{b}/\bar{a}}$ holds by (20), (19) has the form $\tau_1(\bar{a}p_1 + \bar{b}p_2 + Ay) = \varrho\tilde{s}$ and thus we can calculate p_1 and p_2, using the formulas

$$p_2 = (q_2 - v_2)/\bar{\delta}, \quad p_1 = \left(\sqrt{\varrho\bar{a}/\tilde{b}}\,\tilde{s} - Ay - \bar{b}p_2\right) \bigg/ \bar{a}. \tag{29}$$

If $\bar{\delta} = 0$, the choice of q_2 or q_1 is irrelevant; in view of (25), (26) we will suppose from now on that instead of any limited-memory method we use either the shifted DFP method (10) for $\bar{a} \neq 0$, or update $A_+ = A + \varrho\tilde{s}\tilde{s}^{\mathrm{T}}/\tilde{b} - ABss^{\mathrm{T}}BA/\bar{c}$ in the similar form

$$U_+ = U - (1/\bar{c}) \left(\pm\sqrt{\varrho\bar{c}/\tilde{b}}\,\tilde{s} + ABs\right) s^{\mathrm{T}}BU \tag{30}$$

otherwise. In other methods we will suppose $\bar{\delta} \neq 0$, thus $\bar{a}\bar{c} \neq 0$.

We give two methods based on expression (25); some others can be found in [16].

3.1.1. SSBC—simple method based on the shifted Broyden class

Surprisingly, we obtained very good results when we chose simply $q_2 = \hat{w}$, where

$$\hat{w} = \sqrt{\eta\bar{\delta}} \left(\frac{\bar{a}}{\tilde{b}}\,\tilde{s} - Ay\right). \tag{31}$$

Then we have the shifted Broyden update (7) with adding term $-v_2 v_2^{\mathrm{T}}/(\bar{a}\bar{\delta})$.

3.1.2. DSBC—method with direction vector after the shifted Broyden class

Since $d_+ = -H_+ g_+ = -H_+ y - H_+ g = -s + H_+ B d$ by (1) and by $H_+ y = s$ (here we suppose $\varrho = 1$), it suffices to compare value $H_+ B s$, which is

$$\zeta_+ B s + \varrho \frac{\tilde{s}^{\mathrm{T}} B s}{\tilde{b}} \tilde{s} + \frac{q_2^{\mathrm{T}} B s}{\bar{a}\bar{\delta}} q_2 \tag{32}$$

by $v_2^{\mathrm{T}} B s = \bar{\delta}$ for update (25) and

$$\zeta_+ B s + \varrho \frac{\tilde{s}^{\mathrm{T}} B s}{\tilde{b}} \tilde{s} + \frac{1}{\bar{a}} v_2 + \frac{\eta}{\bar{a}} \left(\frac{\bar{a}}{\tilde{b}} \tilde{s}^{\mathrm{T}} B s - \bar{b} \right) \left(\frac{\bar{a}}{\tilde{b}} \tilde{s} - A y \right) \tag{33}$$

for update (7). Comparing (32) with (33), we obtain

$$\frac{q_2^{\mathrm{T}} B s}{\bar{\delta}} q_2 = v_2 + \eta \left(\frac{\bar{a}}{\tilde{b}} \tilde{s}^{\mathrm{T}} B s - \bar{b} \right) \left(\frac{\bar{a}}{\tilde{b}} \tilde{s} - A y \right), \tag{34}$$

which implies

$$\frac{q_2^{\mathrm{T}} B s}{\bar{\delta}} = \pm \sqrt{1 + \frac{\eta}{\bar{\delta}} \left(\frac{\bar{a}}{\tilde{b}} \tilde{s}^{\mathrm{T}} B s - \bar{b} \right)^2}. \tag{35}$$

Combining (34) with (35), we can calculate q_2 for given η (obviously $q_2^{\mathrm{T}} y = 0$) and then p_2 and p_1, using (29).

3.2. Variationally-derived limited-memory methods

Standard VM methods can be obtained by solving a certain variational problem—we find an update with the smallest correction of VM matrix in the sense of some norm (see [10]). Using the product form of the update, we can extend this approach to limited-memory methods to derive a very efficient class of methods. First we give the following general theorem, where the shifted quasi-Newton condition $U_+ U_+^{\mathrm{T}} y = A_+ y = \varrho \tilde{s}$ is equivalently replaced by (the first two conditions imply the third one)

$$U_+^{\mathrm{T}} y = z, \quad U_+ z = \varrho \tilde{s}, \quad z^{\mathrm{T}} z = \varrho \tilde{b}. \tag{36}$$

Theorem 3.1. *Let T be a symmetric positive definite matrix, $z \in \mathcal{R}^m$ and denote $\mathcal{U}$ the set of $N \times m$ matrices. Then the unique solution to*

$$\min\{\varphi(U_+) : U_+ \in \mathcal{U}\} \text{ s.t. } (36), \quad \varphi(U_+) = y^{\mathrm{T}} T y \| T^{-1/2} (U_+ - U) \|_F^2, \tag{37}$$

(*Frobenius matrix norm*) *is*

$$U_+ = U - \frac{T y}{y^{\mathrm{T}} T y} y^{\mathrm{T}} U + \left(\varrho \tilde{s} - U z + \frac{y^{\mathrm{T}} U z}{y^{\mathrm{T}} T y} T y \right) \frac{z^{\mathrm{T}}}{z^{\mathrm{T}} z} \tag{38}$$

and for this solution the value of $\varphi(U_+)$ is

$$\varphi(U_+) = |U^{\mathrm{T}} y - z|^2 + \frac{y^{\mathrm{T}} T y}{z^{\mathrm{T}} z} v^{\mathrm{T}} T^{-1} v, \quad v = \varrho \tilde{s} - U z - \frac{\varrho \tilde{b} - y^{\mathrm{T}} U z}{y^{\mathrm{T}} T y} T y. \tag{39}$$

Proof. Setting $U_+ = (u_1^+, \ldots, u_m^+)$, define Lagrangian function $\mathscr{L} = \mathscr{L}(U_+, e_1, e_2)$ as

$$\begin{aligned}\mathscr{L} &= \frac{1}{2}\,\varphi(U_+) + e_1^{\mathrm{T}}(U_+^{\mathrm{T}} y - z) + e_2^{\mathrm{T}}(U_+ z - \varrho\tilde{s}) \\ &= -\,e_1^{\mathrm{T}} z - \varrho e_2^{\mathrm{T}}\tilde{s} + \sum_{i=1}^{m}\left[\frac{y^{\mathrm{T}}Ty}{2}\,(u_i^+ - u_i)^{\mathrm{T}} T^{-1}(u_i^+ - u_i) + e_{1i}\, y^{\mathrm{T}} u_i^+ + z_i e_2^{\mathrm{T}} u_i^+\right].\end{aligned}$$

A local minimizer U_+ satisfies the equations $\partial\mathscr{L}/\partial u_i^+ = 0$, $i = 1, \ldots, m$, which gives $y^{\mathrm{T}}TyT^{-1}(u_i^+ - u_i) + e_{1i}\,y + z_i e_2 = 0$, $i = 1, \ldots, m$, yielding

$$U_+ = U - \frac{Ty}{y^{\mathrm{T}}Ty}\,e_1^{\mathrm{T}} - \frac{Te_2}{y^{\mathrm{T}}Ty}\,z^{\mathrm{T}}. \tag{40}$$

Using the first condition in (36), we have $e_1 = U^{\mathrm{T}}y - (1 + y^{\mathrm{T}}Te_2/y^{\mathrm{T}}Ty)z$.

Substituting this e_1 to (40), we obtain $U_+ = U - Tyy^{\mathrm{T}}U/y^{\mathrm{T}}Ty + \bar{e}z^{\mathrm{T}}$ with some vector $\bar{e}$. The second condition in (36) yields

$$\bar{e} = \frac{1}{z^{\mathrm{T}}z}\left(\varrho\tilde{s} - Uz + \frac{y^{\mathrm{T}}Uz}{y^{\mathrm{T}}Ty}\,Ty\right) \tag{41}$$

and (38) follows. Matrix U_+ obtained in this way minimizes φ in view of convexity of Frobenius norm. Furthermore, we get

$$\bar{e} - \frac{Ty}{y^{\mathrm{T}}Ty} = \frac{1}{z^{\mathrm{T}}z}\left(\varrho\tilde{s} - Uz - \frac{z^{\mathrm{T}}z - y^{\mathrm{T}}Uz}{y^{\mathrm{T}}Ty}\,Ty\right) = \frac{v}{z^{\mathrm{T}}z} \tag{42}$$

by (36) and (39), thus by (38) and $v^{\mathrm{T}}y = 0$

$$\begin{aligned}\frac{\varphi(U_+)}{y^{\mathrm{T}}Ty} &= \left\| T^{-1/2}\left(\frac{Ty}{y^{\mathrm{T}}Ty}\,y^{\mathrm{T}}U - \bar{e}z^{\mathrm{T}}\right)\right\|_F^2 = \left\| T^{-1/2}\left(\frac{Ty}{y^{\mathrm{T}}Ty}\,(U^{\mathrm{T}}y - z)^{\mathrm{T}} - \frac{v}{z^{\mathrm{T}}z}z^{\mathrm{T}}\right)\right\|_F^2 \\ &= \mathrm{Tr}\left(\frac{(U^{\mathrm{T}}y - z)(U^{\mathrm{T}}y - z)^{\mathrm{T}}}{y^{\mathrm{T}}Ty} + \frac{v^{\mathrm{T}}T^{-1}v}{(z^{\mathrm{T}}z)^2}zz^{\mathrm{T}}\right) = \frac{|U^{\mathrm{T}}y - z|^2}{y^{\mathrm{T}}Ty} + \frac{v^{\mathrm{T}}T^{-1}v}{z^{\mathrm{T}}z}. \qquad \square\end{aligned}$$

The choice of matrix *T*, when vectors *Ty*, $\varrho\tilde{s} - Uz$, are linearly dependent, represents an important special case, since then $v = 0$ (thus the value of $\varphi(U_+)$ reaches its minimum on the set of symmetric positive definite matrices *T*), which implies $\bar{e} = Ty/y^{\mathrm{T}}Ty = (\varrho\tilde{s} - Uz)/(\varrho\tilde{b} - y^{\mathrm{T}}Uz)$ by (42) and in view of (41), update (38) can be written in the form

$$U_+ = U - \frac{\varrho\tilde{s} - Uz}{\varrho\tilde{b} - y^{\mathrm{T}}Uz}(U^{\mathrm{T}}y - z)^{\mathrm{T}}. \tag{43}$$

General form of variationally-derived update (38) can be rewritten, using (36):

$$U_+ = \frac{\tilde{s}z^{\mathrm{T}}}{\tilde{b}} + \left(I - \frac{Tyy^{\mathrm{T}}}{y^{\mathrm{T}}Ty}\right)U\left(I - \frac{zz^{\mathrm{T}}}{z^{\mathrm{T}}z}\right). \tag{44}$$

Since $z^{\mathrm{T}}(I - zz^{\mathrm{T}}/z^{\mathrm{T}}z) = 0$ and $(I - zz^{\mathrm{T}}/z^{\mathrm{T}}z)^2 = I - zz^{\mathrm{T}}/z^{\mathrm{T}}z$, this yields

$$A_+ = \varrho\,\frac{\tilde{s}\tilde{s}^{\mathrm{T}}}{\tilde{b}} + \left(I - \frac{Tyy^{\mathrm{T}}}{y^{\mathrm{T}}Ty}\right) U \left(I - \frac{zz^{\mathrm{T}}}{z^{\mathrm{T}}z}\right) U^{\mathrm{T}} \left(I - \frac{yy^{\mathrm{T}}T}{y^{\mathrm{T}}Ty}\right) \tag{45}$$

by $A_+ = U_+U_+^{\mathrm{T}}$. This expression, which can be easily compared with the quasi-product form (9) of the shifted Broyden class update (which we get for $Ty = (\sqrt{\eta}/\tilde{b})\tilde{s} + ((1-\sqrt{\eta})/\bar{a})Ay$ and Uz proportional to Ty), shows the meaning of parameters z, Ty.

Using Theorem 3.1 for the standard Broyden class (see [10]), we can easily derive the new product form of these updates. To do it, we set $H = SS^{\mathrm{T}}$ and replace $U, \tilde{s}, \tilde{b}$ by S, s, b. Then for $Ty = (\sqrt{\eta}/b)s + ((1-\sqrt{\eta})/a)Hy$, $z = \alpha b S^{\mathrm{T}}BTy$, $\eta \geqslant 0$, $\alpha \in \mathscr{R}$, update (38) will be replaced by $S_+ = S - Tyy^{\mathrm{T}}S + (\varrho/z^{\mathrm{T}}z)sz^{\mathrm{T}}$ by proportionality of Sz, Ty and we have

$$S_+ = S - Ty(S^{\mathrm{T}}y)^{\mathrm{T}} + \alpha s(S^{\mathrm{T}}BTy)^{\mathrm{T}} \tag{46}$$

by $z^{\mathrm{T}}z = \varrho b$, which is the product form of updates from the Broyden class for $\eta \geqslant 0$:

Theorem 3.2. *Every update* (46) *with* $Ty = (\sqrt{\eta}/b)s + ((1-\sqrt{\eta})/a)Hy$, $\eta \geqslant 0$, $\alpha^2 = \varrho ab/[b^2 + \eta(ac - b^2)]$ *belongs to the Broyden class generated by the parameter* η.

Proof. We can utilize Lemma 2.2 in [17] or use straightforward arrangements and compare corresponding terms. □

The following two methods are based on this comparison with the BFGS update ($\eta = 1$). Note that neither update (48) nor (49) need not calculate vector Ay. These methods were implemented in subroutine PLIP, see [9], which can be downloaded from www.cs.cas.cz/~luksan/subroutines.html.

3.2.1. VAR1—type 1 variationally-derived method

By analogy with the product form of the BFGS update ($\eta = 1$, $z = \alpha S^{\mathrm{T}}Bs$), we set

$$z = \vartheta U^{\mathrm{T}}Bs, \quad \vartheta = \pm\sqrt{\varrho\tilde{b}/\bar{c}}, \tag{47}$$

by $z^{\mathrm{T}}z = \varrho\tilde{b}$. Then (43) gives

$$U_+ = U - \frac{\varrho\tilde{s} - \vartheta ABs}{\varrho\tilde{b} - \vartheta\bar{b}}(y - \vartheta Bs)^{\mathrm{T}}U, \tag{48}$$

which gives the best results for the choice $\operatorname{sgn}(\vartheta\bar{b}) = -1$ (compare with Theorem 4.5).

3.2.2. VAR2—type 2 variationally-derived method

With z given by (47) and with the simple choice $Ty = \tilde{s}$, (38) leads to type 2 method

$$U_+ = U - \frac{\tilde{s}}{\tilde{b}}y^{\mathrm{T}}U + \left[\left(\frac{\varrho}{\vartheta} + \frac{\bar{b}}{\tilde{b}}\right)\tilde{s} - ABs\right]\frac{s^{\mathrm{T}}BU}{\bar{c}}. \tag{49}$$

Efficiency of both these methods significantly depends on the value of the correction parameter ϱ. The recommended value is $\varrho^{(1)} = \zeta/(\zeta + \zeta_+)$, which is suitable for the most of problems. Very good results were also obtained with the choices: $\varrho^{(2)} = \sqrt[4]{\mu^2 \varrho^{(1)}/2}$, $\varrho^{(3)} = \nu$, $\varrho^{(4)} = \sqrt{\nu\varepsilon}$, where $\nu = \mu/(1-\mu)$, μ is a relative shift parameter and $\varepsilon = \sqrt{\zeta\hat{a}/a}$ is the damping factor of μ, see Section 2.2. Note that for choice $\varrho = \nu$ equality $y^{\mathrm{T}} A_+ y = \zeta_+ y^{\mathrm{T}} y$ holds by (6) and (11), i.e. this value balances the both parts of $y^{\mathrm{T}} H_+ y = \zeta_+ y^{\mathrm{T}} y + y^{\mathrm{T}} A_+ y$.

3.3. Computational experiments

Our new limited-memory VM methods were tested, using the collection of relatively difficult problems with optional dimension chosen from [12] (Test 28, some problems are dense) and collection of problems for general sparse and partially separable unconstrained optimization from [11] (Test 14, usually well-conditioned problems). We have used $m=10, 20$ for $N=1000$ and $m=5, 10$ for $N=5000$, the final precision $\|g(x^\star)\|_\infty \leqslant 10^{-6}$, $\eta=1$ for the corresponding shifted Broyden class (methods SSBC and DSBC) and the choice of the shift parameter μ after (15) (the recommended value). For starting iterates we use the shifted BFGS method.

Results of our experiments are given in three tables, where NIT is the total number of iterations (over all problems), NFV the total number of function and also gradient evaluations, 'Fail' denotes the number of problems which were not solved successfully (usually NFV reached its limit) and 'Time' is the total computational time. The first four rows of tables give results for methods SSBC, DSBC, VAR1 and VAR2. In case variationally-derived methods we used $\varrho = \varrho^{(2)}$ for method VAR1 and $\varrho = \varrho^{(1)}$ for method VAR2 in Table 3 (see Section 2.3) and $\varrho = \varrho^{(1)}$ in Tables 4 and 5.

For comparison, the last four rows contain results for the following limited-memory methods: LBFGS—the Nocedal method based on the Strang formula, see [13], BNS—the method after [3], RH—the reduced-Hessian method described in [7] and CG—the conjugate gradient method (Hestenes and Stiefel version), see [6]; this method often stopped before the requested precision was achieved. Note that methods BNS and LBFGS store $2m$ vectors while method CG stores no additional vectors. From our numerical experiments we may state that variationally derived methods VAR1 and especially VAR2 are usually better than methods SSBC and DSBC.

For a better demonstration of both the efficiency and the reliability, we compare selected optimization methods by using performance profiles introduced in [4]. The performance profile $\pi_M(\tau)$ is defined by

Table 3
(Test 28, $N = 1000$, 80 problems)

Method	$m = 10$				$m = 20$			
	NIT	NFV	Fail	Time	NIT	NFV	Fail	Time
SSBC	97991	100990	—	46.3	95012	98314	—	62.2
DSBC	105976	109096	—	51.6	103383	106328	—	66.2
VAR1	95495	99541	—	42.6	95327	98775	—	51.8
VAR2	91585	95304	—	41.8	84671	87964	—	48.6
LBFGS	92800	98921	—	37.6	86899	92294	—	44.7
BNS	91234	95532	—	40.9	93397	97704	—	56.7
RH	91160	113314	—	40.4	101251	122853	—	56.1
CG	108770	223626	4	59.6				

Table 4
(Test 14, $N = 1000$, 22 problems)

Method	$m = 10$				$m = 20$			
	NIT	NFV	Fail	Time	NIT	NFV	Fail	Time
SSBC	20095	20312	—	12.22	17936	18142	—	12.66
DSBC	21874	22150	—	13.33	18428	18677	—	13.49
VAR1	19260	19660	—	10.42	17162	17472	—	10.77
VAR2	18430	18693	—	10.20	16499	16735	—	11.00
LBFGS	20337	21383	—	11.00	18578	19590	—	11.40
BNS	21017	22097	—	12.36	19625	20613	—	14.41
RH	21892	33442	—	18.63	21526	33134	—	24.16
CG	20003	40034	—	12.12				

Table 5
(Test 14, $N = 5000$, 20 problems)

Method	$m = 5$				$m = 10$			
	NIT	NFV	Fail	Time	NIT	NFV	Fail	Time
SSBC	109342	109917	2	6:11.1	88063	88468	—	6:04.7
DSBC	104763	105646	1	5:27.3	93295	93929	—	6:13.4
VAR1	97057	98888	—	4:43.0	68561	69811	—	3:57.6
VAR2	87713	89500	—	4:21.8	67360	68637	—	3:54.1
LBFGS	106345	109387	2	4:38.1	82311	84446	—	4:27.7
BNS	104569	107467	2	5:07.1	85681	87827	—	4:55.3
RH	97037	155691	4	6:58.6	86402	137572	2	6:24.1
CG	57056	192346	4	7:49.3				

the formula

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

with $\tau \geqslant 0$, where $\tau_{P,M}$ is the performance ratio of the time (or the number of function evaluations) required to solve problem P by method M to the lowest time (or the number of function evaluations) required to solve problem P. The ratio $\tau_{P,M}$ is set to infinity (or some large number) if method M fails to solve problem P. The value of $\pi_M(\tau)$ at $\tau = 0$ gives the percentage of test problems for which the method M is the best and the value for τ large enough is the percentage of test problems that method M can solve. The relative efficiency and reliability of each method can be directly seen from the performance profiles: the higher is the particular curve the better is the corresponding method. The following figures (Figs. 1–3) reveal the performance profiles for methods VAR2, LBFGS and RH graphically. These figures are based on results used in the left parts of the previous tables.

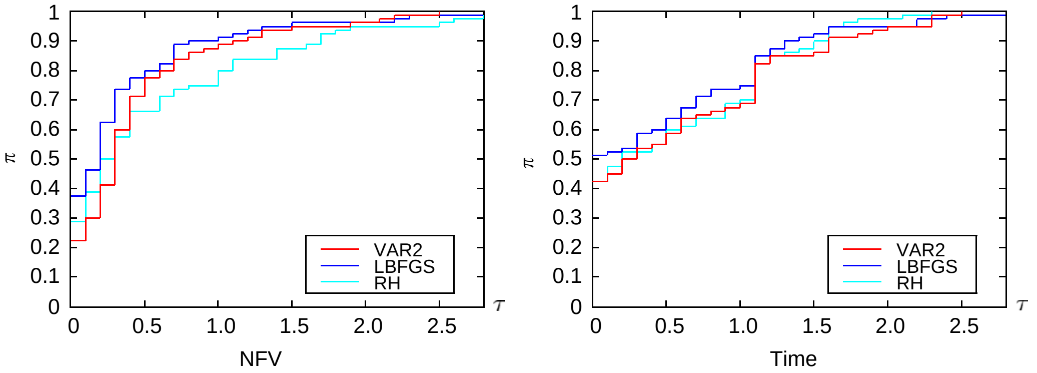

Fig. 1. (Test 28, $N = 1000$, $m = 10$, 80 problems).

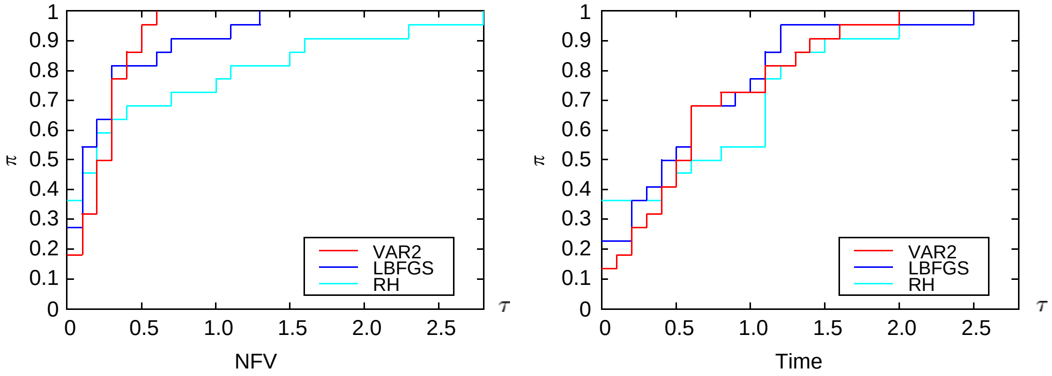

Fig. 2. (Test 14, $N = 1000$, $m = 10$, 22 problems).

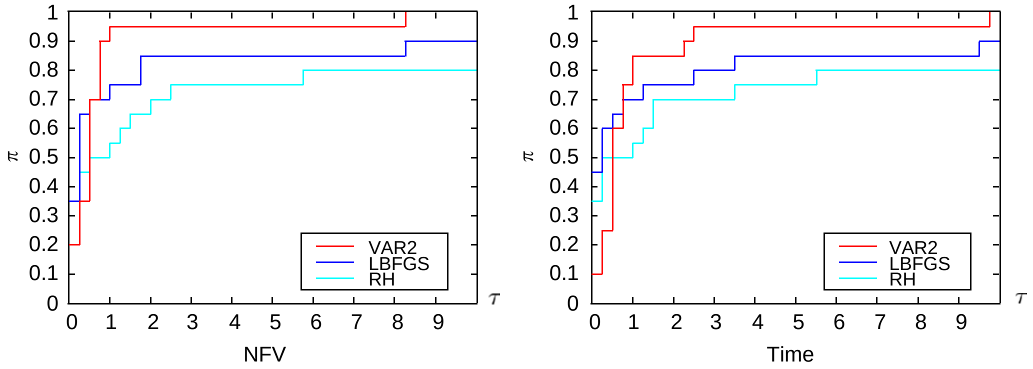

Fig. 3. (Test 14, $N = 5000$, $m = 5$, 20 problems).

4. Global convergence

In this section we establish global convergence of methods from the shifted Broyden class with $\eta \in [0, 1]$ and our limited-memory methods for f uniformly convex. At the end we describe a simple way allowing us to assure global convergence in the nonconvex case.

Assumption 4.1. The objective function $f : \mathscr{R}^N \to \mathscr{R}$ is bounded from below and uniformly convex with bounded second-order derivatives (i.e., $0 < \underline{G} \leqslant \underline{\lambda}(G(x)) \leqslant \overline{\lambda}(G(x)) \leqslant \overline{G} < \infty$, $x \in \mathscr{R}^N$, where $\underline{\lambda}(G(x))$ and $\overline{\lambda}(G(x))$ are the lowest and the greatest eigenvalues of the Hessian matrix $G(x)$).

Assumption 4.2. Parameters ϱ_k and μ_k of the shifted VM method are uniformly positive and bounded, in the sense that $0 < \underline{\varrho} \leqslant \varrho_k \leqslant \overline{\varrho}$, $0 < \underline{\mu} \leqslant \mu_k \leqslant \overline{\mu} < 1$, $k \geqslant 1$.

Lemma 4.1. *Let the objective function satisfy Assumption* 4.1 *and parameter* μ *satisfy Assumption* 4.2. *Then* $\hat{a}/b \in [\underline{G}, \overline{G}]$ *and* $b/|\tilde{s}|^2 > b/|s|^2 \geqslant \underline{G}$.

Proof. Setting $G_I = \int_0^1 G(x + \xi s)\,\mathrm{d}\xi$, $q = G_I^{1/2} s$, we obtain $y = g_+ - g = G_I s$ and thus

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

by Assumption 4.1. Similarly, $b/|s|^2 = s^{\mathrm{T}} G_I s / s^{\mathrm{T}} s = \int_0^1 s^{\mathrm{T}} G(x + \xi s) s / s^{\mathrm{T}} s\,\mathrm{d}\xi \geqslant \underline{G}$ and $|\tilde{s}|^2 = |s - (\mu b/\hat{a}) y|^2 = |s|^2 - \mu(2 - \mu) b^2/\hat{a} < |s|^2$ by (6), (11) and Assumption 4.2. □

4.1. Shifted Broyden class and modified shifted DFP method

Theorem 4.1. *Consider any shifted variable metric method satisfying* (5) *and* (6). *Let the objective function satisfy Assumption* 4.1 *and parameter* μ *satisfy Assumption* 4.2, *with the line-search method fulfilling* (1) *and* (2). *If there is a constant* $0 < C < \infty$ *that*

$$\operatorname{Tr} A_{k+1} \leqslant \operatorname{Tr} A_k + C, \quad k \geqslant 1, \tag{50}$$

then (3) *holds.*

Proof. Since $\hat{a}/b \in [\underline{G}, \overline{G}]$ by Lemma 4.1, Assumption 4.2 implies $\zeta_{k+1} \in [\underline{\zeta}, \overline{\zeta}]$, $k \geqslant 1$, by (11), where $\underline{\zeta} = \underline{\mu}/\overline{G}$ and $\overline{\zeta} = \overline{\mu}/\underline{G}$. Using (50), one has

$$\|H_{k+1}\| \leqslant \zeta_{k+1} + \|A_{k+1}\| \leqslant \overline{\zeta} + \operatorname{Tr} A_{k+1} \leqslant \overline{\zeta} + \operatorname{Tr} A_1 + Ck \leqslant \tilde{C}(k + 1), \quad k \geqslant 1,$$

where $\tilde{C} = \max(\overline{\zeta} + \operatorname{Tr} A_1, C)$. By (1) and (5), this gives

$$\cos^2 \theta_k \triangleq \frac{(g_k^{\mathrm{T}} d_k)^2}{g_k^{\mathrm{T}} g_k d_k^{\mathrm{T}} d_k} = \frac{g_k^{\mathrm{T}} (\zeta_k I + A_k) g_k}{g_k^{\mathrm{T}} g_k} \frac{g_k^{\mathrm{T}} H_k g_k}{g_k^{\mathrm{T}} H_k^2 g_k} \geqslant \zeta_k \frac{1}{\|H_k\|} \geqslant \frac{\underline{\zeta}}{\tilde{C} k}, \quad k \geqslant 1.$$

Thus $\sum_{k=1}^{\infty} \cos^2 \theta_k = \infty$ and (3) follows from Theorem 1.1. □

Corollary 4.1. *Let the objective function satisfy Assumption* 4.1 *and parameters* ϱ *and* μ *satisfy Assumption* 4.2. *Suppose that the line-search method fulfils* (1) *and* (2). *Then* (3) *holds for the shifted variable metric method* (7) *with* $\eta \in [0, \varrho/(\varrho + \bar{a}/\tilde{b})]$.

Proof. Consider update

$$A_+ = A + 2\varrho \frac{\tilde{s}\tilde{s}^{\mathrm{T}}}{\tilde{b}} - \frac{(\varrho\tilde{s} + Ay)(\varrho\tilde{s} + Ay)^{\mathrm{T}}}{\bar{a} + \tilde{b}\varrho}, \tag{51}$$

which belongs to the shifted Broyden class (7) with $\eta = \varrho/(\varrho + \bar{a}/\tilde{b})$ and represents the shifted analogy of Hoshino self-dual method, see [10]. Since

$$\mathrm{Tr}\, A_+ \leqslant \mathrm{Tr}\, A + 2\varrho|\tilde{s}|^2/\tilde{b} \leqslant \mathrm{Tr}\, A + 2\overline{\varrho}/[(1-\overline{\mu})\underline{G}] \tag{52}$$

by Lemma 4.1 and Assumption 4.2, method (51) is globally convergent by Theorem 4.1. By (7), (52) obviously holds also for methods from the shifted Broyden class with $\eta \leqslant \varrho/(\varrho + \bar{a}/\tilde{b})$. □

Now we establish global convergence of all methods from the shifted Broyden class with $\eta \in [0, 1]$, using additional assumption $\mu^2 \leqslant \zeta\hat{a}/a$, which corresponds to the choice of coefficient ε for the shift parameter μ (see Section 2.2) and which is satisfied for μ given by (14). Note that this assumption can be significantly weakened, see Lemma 4.4. Denote $\tilde{H}_+ = \zeta I + A_+$. The following lemma plays basic role.

Lemma 4.2. *Consider the shifted variable metric method* (7) *with* $\eta \in [0, 1]$. *Then*

$$\frac{\det \tilde{H}_+}{\det H} \leqslant \frac{\tilde{s}^{\mathrm{T}} B\tilde{s}}{\tilde{b}} \left(\varrho + \frac{\zeta\hat{a}}{\tilde{b}}\right). \tag{53}$$

Proof. It suffices to prove the desired inequality for $\eta = 1$ by (7) and the identity $\det(\tilde{H}_+ - uu^{\mathrm{T}}) = (1 - u^{\mathrm{T}}\tilde{H}_+^{-1}u) \det \tilde{H}_+$. The shifted BFGS update (8) can be rewritten $A_+ = A + [(\omega\tilde{s} - Ay)(\omega\tilde{s} - Ay)^{\mathrm{T}} - Ayy^{\mathrm{T}}A]/(\tilde{b}\omega)$, where $\omega = \varrho + \bar{a}/\tilde{b}$, or

$$\tilde{H}_+ = H^{1/2}\left(I + \frac{B^{1/2}(\omega\tilde{s} - Ay)(\omega\tilde{s} - Ay)^{\mathrm{T}}B^{1/2} - B^{1/2}Ayy^{\mathrm{T}}AB^{1/2}}{\tilde{b}\omega}\right)H^{1/2}.$$

Since

$$\begin{aligned}\det(I + (u-v)(u-v)^{\mathrm{T}} - vv^{\mathrm{T}}) &= (1 + |u-v|^2)(1 - |v|^2) + ((u-v)^{\mathrm{T}}v)^2 \\ &= |u|^2 + (1 - u^{\mathrm{T}}v)^2 - |u|^2|v|^2,\end{aligned}$$

we obtain

$$\frac{\det \tilde{H}_+}{\det H} = \omega\frac{\tilde{s}^{\mathrm{T}}B\tilde{s}}{\tilde{b}} + \left(1 - \frac{\tilde{s}^{\mathrm{T}}BAy}{\tilde{b}}\right)^2 - \frac{\tilde{s}^{\mathrm{T}}B\tilde{s}y^{\mathrm{T}}ABAy}{\tilde{b}^2}.$$

Observing that $\tilde{s}^{\mathrm{T}}BAy=\tilde{b}-\zeta\tilde{s}^{\mathrm{T}}By$ and $y^{\mathrm{T}}ABAy=\bar{a}-\zeta\hat{a}+\zeta^2y^{\mathrm{T}}By$, we find

$$\begin{aligned}\frac{\det\tilde{H}_+}{\det H}&=\omega\,\frac{\tilde{s}^{\mathrm{T}}B\tilde{s}}{\tilde{b}}+\frac{\zeta^2(\tilde{s}^{\mathrm{T}}By)^2}{\tilde{b}^2}-\frac{\tilde{s}^{\mathrm{T}}B\tilde{s}\,y^{\mathrm{T}}ABAy}{\tilde{b}^2}\\&=\left(\varrho+\frac{\zeta\hat{a}}{\tilde{b}}\right)\frac{\tilde{s}^{\mathrm{T}}B\tilde{s}}{\tilde{b}}+\zeta^2\,\frac{(\tilde{s}^{\mathrm{T}}By)^2-\tilde{s}^{\mathrm{T}}B\tilde{s}\,y^{\mathrm{T}}By}{\tilde{b}^2}\leqslant\left(\varrho+\frac{\zeta\hat{a}}{\tilde{b}}\right)\frac{\tilde{s}^{\mathrm{T}}B\tilde{s}}{\tilde{b}}\end{aligned}$$

by the Schwarz inequality. □

Lemma 4.3. *Consider any shifted variable metric method satisfying* (5) *and* (6). *Then*

$$\det H_+/\det\tilde{H}_+<(1+\zeta_+/\zeta)^N. \tag{54}$$

Proof. Denoting $\tilde{\lambda}_1,\ldots,\tilde{\lambda}_N$ the eigenvalues of $\tilde{H}_+$, we have $\tilde{\lambda}_i\geqslant\zeta,\ \ i=1,\ldots,N$ in view of $\tilde{H}_+=\zeta I+A_+$. Since $H_+=\tilde{H}_++(\zeta_+-\zeta)I$, we obtain

$$\det H_+/\det\tilde{H}_+=(1+(\zeta_+-\zeta)/\tilde{\lambda}_1)\cdots(1+(\zeta_+-\zeta)/\tilde{\lambda}_N)<(1+\zeta_+/\zeta)^N. \qquad \square$$

Lemma 4.4. *Consider any shifted variable metric method satisfying* (5) *and* (6). *If there is a constant C that* $\mu^2\leqslant C\zeta\hat{a}c/b^2$, *e.g., if* $\mu^2\leqslant C\zeta\hat{a}/a$, *then* $\tilde{s}^{\mathrm{T}}B\tilde{s}\leqslant c(1+\sqrt{C})^2$.

Proof. We have $c/b^2\geqslant 1/a$ by the Schwarz inequality. Assumption $\mu^2\leqslant C\zeta\hat{a}c/b^2$ implies $\zeta_+^2=\mu^2(b/\hat{a})^2\leqslant C\zeta c/\hat{a}$. Observing that $\zeta y^{\mathrm{T}}By/y^{\mathrm{T}}y\leqslant\zeta\|B\|\leqslant 1$ by (5), we have $\zeta_+^2y^{\mathrm{T}}By\leqslant cC\zeta y^{\mathrm{T}}By/\hat{a}\leqslant cC$. Since $\tilde{s}=s-\zeta_+y$, we get by the Schwarz inequality

$$\tilde{s}^{\mathrm{T}}B\tilde{s}=c-2\zeta_+s^{\mathrm{T}}By+\zeta_+^2y^{\mathrm{T}}By\leqslant\left(\sqrt{c}+\zeta_+\sqrt{y^{\mathrm{T}}By}\right)^2\leqslant c\left(1+\sqrt{C}\right)^2. \qquad \square$$

Lemma 4.5. *Consider any shifted variable metric method satisfying* (5) *and* (6) *and Assumption* 4.2. *Let the objective function satisfy Assumption* 4.1. *Then* $\zeta_+\in[\underline{\zeta},\overline{\zeta}]\triangleq[\underline{\mu}/\overline{G},\overline{\mu}/\underline{G}]$. *Moreover, if* $\mu=\underline{\mu}$, *then* $\zeta_+/\zeta\leqslant\overline{G}/\underline{G}$ *and* $\tilde{s}^{\mathrm{T}}B\tilde{s}\leqslant 2c+2\underline{\mu}b\overline{G}/\underline{G}$.

Proof. Since $\hat{a}/b\in[\underline{G},\overline{G}]$ by Lemma 4.1 and $\zeta_+=\mu b/\hat{a}$ by (11), we deduce $\zeta_+\in[\underline{\mu}/\overline{G},\overline{\mu}/\underline{G}]$. Let $\mu=\underline{\mu}$. Then we have $\zeta_+/\zeta\leqslant\zeta_+\overline{G}/\underline{\mu}\leqslant\overline{G}/\underline{G}$. Using inequalities $\tilde{s}^{\mathrm{T}}B\tilde{s}\leqslant(\sqrt{c}+\zeta_+\sqrt{y^{\mathrm{T}}By})^2$and $\zeta y^{\mathrm{T}}By\leqslant y^{\mathrm{T}}y$, see the proof of Lemma 4.4, we obtain

$$\tilde{s}^{\mathrm{T}}B\tilde{s}\leqslant\left(\sqrt{c}+\zeta_+\sqrt{y^{\mathrm{T}}By}\right)^2\leqslant 2(c+\zeta_+^2y^{\mathrm{T}}By)\leqslant 2c+2\zeta_+^2\hat{a}/\zeta\leqslant 2c+2\underline{\mu}b\overline{G}/\underline{G}. \qquad \square$$

Theorem 4.2. *Consider the shifted variable metric method* (7) *satisfying Assumption* 4.2 *with* $\underline{\mu}$ *sufficiently small and suppose that the line-search method fulfils* (1) *and* (2). *Let the objective function satisfy Assumption* 4.1. *If* $\eta\in[0,1]$ *and* $\mu^2\leqslant\zeta\hat{a}/a$ *or* $\mu=\underline{\mu}$ *(e.g., if* $\underline{\mu}^2>\zeta\hat{a}/a$*), then* (3) *holds.*

Proof. Combining Lemmas 4.2 and 4.3, we find $\det H_+/\det H < \chi\tilde{s}^{\mathrm{T}}B\tilde{s}/\tilde{b}$, where $\chi = (\varrho + \zeta\hat{a}/\tilde{b})(1 + \zeta_+/\zeta)$. Observing that $\det H \geqslant \zeta^N \geqslant \underline{\zeta}^N$ by (5), we get

$$C_1 \triangleq \frac{\underline{\zeta}^N}{\det H_2} \leqslant \frac{\det H_{k+2}}{\det H_2} = \prod_{i=2}^{k+1} \frac{\det H_{i+1}}{\det H_i} < \prod_{i=2}^{k+1} \chi_i \, \frac{\tilde{s}_i^{\mathrm{T}} B_i \tilde{s}_i}{\tilde{b}_i} \leqslant \left(\frac{1}{k} \sum_{i=2}^{k+1} \chi_i \, \frac{\tilde{s}_i^{\mathrm{T}} B_i \tilde{s}_i}{\tilde{b}_i} \right)^k, \tag{55}$$

$k \geqslant 1$, $C_1 > 0$. Since always $\tilde{s}^{\mathrm{T}}B\tilde{s} \leqslant 4c + 2\underline{\mu}b\overline{G}/\underline{G}$ by Lemma 4.4 with $C = 1$ and Lemma 4.5, $\tilde{b} = b(1-\mu)$ by (11) and since $\chi/(1-\mu) \leqslant C_2(1+\zeta_+/\zeta)$ with $C_2 = (\overline{\varrho} + \overline{\zeta G})/(1-\overline{\mu})^2$ by Lemma 4.1 and Lemma 4.5, we obtain from (55)

$$kC_1^{1/k} < \sum_{i=2}^{k+1} \frac{\tilde{s}_i^{\mathrm{T}} B_i \tilde{s}_i}{b_i} \, \frac{\chi_i}{1-\mu_i} \leqslant 4C_2 \sum_{i=2}^{k+1} \frac{c_i}{b_i} \left(1 + \frac{\zeta_{i+1}}{\zeta_i}\right) + 2C_2\underline{\mu} \frac{\overline{G}}{\underline{G}} \sum_{\substack{i=2 \\ \mu_i = \underline{\mu}}}^{k+1} \left(1 + \frac{\zeta_{i+1}}{\zeta_i}\right),$$

$k \geqslant 1$. Using Lemma 4.5, we get

$$4C_2 \left(1 + \frac{\overline{\zeta}}{\underline{\zeta}}\right) \sum_{i=2}^{k+1} \frac{c_i}{b_i} > k[C_1^{1/k} - \underline{\mu}C_3], \quad C_3 = 2C_2 \frac{\overline{G}}{\underline{G}} \left(1 + \frac{\overline{G}}{\underline{G}}\right), \tag{56}$$

$k \geqslant 1$. Let $\underline{\mu}$ be chosen in such a way that $\underline{\mu} < 1/C_3$. Observing that $C_1^{1/k} \xrightarrow{k} 1$, (56) implies $\sum_{i=2}^{k+1} c_i/b_i \xrightarrow{k} \infty$. Since $g^{\mathrm{T}}Hg \geqslant \zeta g^{\mathrm{T}}g \geqslant \underline{\zeta}g^{\mathrm{T}}g$ by (5), we obtain for $k \geqslant 1$

$$\sum_{i=2}^{k+1} \cos^2\theta_i \triangleq \sum_{i=2}^{k+1} \frac{(g_i^{\mathrm{T}}d_i)^2}{g_i^{\mathrm{T}}g_i d_i^{\mathrm{T}}d_i} = \sum_{i=2}^{k+1} \frac{g_i^{\mathrm{T}}d_i}{g_i^{\mathrm{T}}g_i} \, \frac{t_i g_i^{\mathrm{T}}s_i}{s_i^{\mathrm{T}}s_i} = \sum_{i=2}^{k+1} \frac{g_i^{\mathrm{T}}H_i g_i}{g_i^{\mathrm{T}}g_i} \, \frac{b_i}{s_i^{\mathrm{T}}s_i} \, \frac{c_i}{b_i} \geqslant \underline{\zeta}\underline{G} \sum_{i=2}^{k+1} \frac{c_i}{b_i}$$

by (1) and Lemma 4.1. Thus $\sum_{i=1}^{\infty} \cos^2\theta_i = \infty$ and (3) follows from Theorem 1.1. □

We recall that assumption $\mu^2 \leqslant \zeta\hat{a}/a$ corresponds to the choice of coefficient ε for the shift parameter μ (see Section 2.2).

The bound $1/C_3$ does not give a realistic estimate for $\underline{\mu}$, e.g., since the number of cases when $\underline{\mu}^2 > \zeta\hat{a}/a$ can be negligible. We tested various choices of $\underline{\mu}$ and found that methods in Section 2 give the best results with the choice (14) without any corrections (with the exception of initial iterations), while in case of methods in Section 3 (their global convergence properties are also based on Theorem 4.2) better results were obtained with corrections (15) in every iteration, i.e. with $\underline{\mu} = 0.2$.

4.2. Limited-memory methods

We utilize expressions (25) and (27) obtained in Section 3.1. The following basic assertion holds.

Theorem 4.3. *Denote* $\hat{w} = \sqrt{\eta\bar{\delta}}((\bar{a}/\tilde{b})\tilde{s} - Ay)$ *and consider the shifted variable metric method* (25) *with* $q_2 = \alpha\hat{w} + \beta v_2$ (*or method* (27) *with* $q_1\sqrt{\bar{a}/\bar{c}} = \alpha\hat{w} + \beta v_2$), *satisfying Assumption* 4.2 *with* $\underline{\mu}$ *sufficiently small and suppose that the line-search method fulfils* (1) *and* (2). *Let the objective function satisfy Assumption* 4.1. *If* $\alpha^2 + \beta^2 \leqslant 1$, $\eta \in [0, 1]$ *and* $\mu^2 \leqslant \zeta\hat{a}/a$ *or* $\mu = \underline{\mu}$, *then* (3) *holds.*

Proof. If $\bar{\delta} \neq 0$, we can obviously restrict to update (25) and write by assumption

$$q_2 q_2^{\mathrm{T}} - v_2 v_2^{\mathrm{T}} = \alpha^2 \hat{w}\hat{w}^{\mathrm{T}} + \alpha\beta\hat{w}v_2^{\mathrm{T}} + \alpha\beta v_2\hat{w}^{\mathrm{T}} + (\beta^2 - 1)v_2 v_2^{\mathrm{T}}. \tag{57}$$

First suppose that $\beta^2 < 1$. Denoting $\eta' = \eta\alpha^2/(1-\beta^2) \leqslant \eta \leqslant 1$, (57) yields

$$\frac{q_2 q_2^{\mathrm{T}} - v_2 v_2^{\mathrm{T}}}{\bar{a}\bar{\delta}} = \frac{\eta'}{\bar{a}}\left(\frac{\bar{a}}{\tilde{b}}\tilde{s} - Ay\right)\left(\frac{\bar{a}}{\tilde{b}}\tilde{s} - Ay\right)^{\mathrm{T}} - uu^{\mathrm{T}}, \quad u = \frac{(1-\beta^2)v_2 - \alpha\beta\hat{w}}{\sqrt{\bar{a}\bar{\delta}(1-\beta^2)}},$$

by $\bar{a}\bar{\delta}uu^{\mathrm{T}} = (1-\beta^2)v_2v_2^{\mathrm{T}} - \alpha\beta v_2\hat{w}^{\mathrm{T}} - \alpha\beta\hat{w}v_2^{\mathrm{T}} + (\eta'/\eta)\beta^2\hat{w}\hat{w}^{\mathrm{T}}$. Therefore (25) represents update (7) with adding term $-uu^{\mathrm{T}}$. Without this adding term, this update satisfies assumptions of Lemma 4.2 and inequality (53) holds by identity $\det(\tilde{H}_+ - uu^{\mathrm{T}}) = (1 - u^{\mathrm{T}}\tilde{H}_+^{-1}u)\det \tilde{H}_+$. If $\beta^2 = 1$, condition $\alpha^2 + \beta^2 \leqslant 1$ implies $\alpha = 0$ and (25) represents the shifted DFP method, which also satisfies assumptions of Lemma 4.2. Thus (53) holds and the desired result follows as in the proof of Theorem 4.2.

Obviously, the case $\bar{\delta} = 0$ does not violate global convergency, since we use either the shifted DFP method (see Section 3.1) for $\bar{a} \neq 0$, or update $A_+ = A + \varrho\tilde{s}\tilde{s}^{\mathrm{T}}/\tilde{b} - ABss^{\mathrm{T}}BA/\bar{c}$ i.e. the shifted DFP method (8) with adding term $-ABss^{\mathrm{T}}BA/\bar{c}$ otherwise. This is also relevant to all methods in this section. □

Corollary 4.2. *Let the objective function satisfy Assumption* 4.1 *and* $\mu^2 \leqslant \zeta\hat{a}/a$ *or* $\mu = \underline{\mu}$ *and suppose that the line-search method fulfils* (1) *and* (2). *For methods SSBC and DSBC* (*see Section* $\overline{3}$.1), *satisfying Assumption* 4.2 *with* $\underline{\mu}$ *sufficiently small*, (3) *holds*.

Proof. We have $\alpha = 1$, $\beta = 0$ for the first method. For the second method, we obtain

$$\alpha = \pm\hat{w}^{\mathrm{T}}Bs \Big/ \sqrt{\bar{\delta}^2 + (\hat{w}^{\mathrm{T}}Bs)^2}\,, \quad \beta = \pm\bar{\delta} \Big/ \sqrt{\bar{\delta}^2 + (\hat{w}^{\mathrm{T}}Bs)^2}\,,$$

by (34) and (35), thus $\alpha^2 + \beta^2 = 1$ for both these methods and we use Theorem 4.3. □

Now we concentrate on update (38) with the choice (47), which is type 2 method with $p_1 = -Ty/y^{\mathrm{T}}Ty$. Thus $p_1^{\mathrm{T}}y = -1$, yielding $q_1^{\mathrm{T}}y = -\bar{\delta} + v_1^{\mathrm{T}}y = 0$. Therefore we can express this update in the form (27) and use the following theorem.

Theorem 4.4. *Let* $\eta > 0$. *Consider update* (38) *with the choice* (47) *and with*

$$Ty = \tilde{s} + \beta_1 ABs + \beta_2 Ay. \tag{58}$$

If

$$(\bar{a}\beta_2 + \tilde{b})^2 \geqslant \bar{a}\bar{c}\beta_1^2 + \tilde{b}^2/\eta \tag{59}$$

holds, then the assumption $\alpha^2 + \beta^2 \leqslant 1$ *of Theorem* 4.3 *is satisfied*.

Proof. From $p_1 = -Ty/y^{\mathrm{T}}Ty$ and (58) we obtain

$$
\begin{aligned}
q_1 &= \bar{\delta} p_1 + v_1 = -\bar{\delta}\, \frac{\tilde{s} - (\tilde{b}/\bar{a})Ay + \beta_1 ABs + (\beta_2 + \tilde{b}/\bar{a})Ay}{\tilde{b} + \bar{b}\beta_1 + \bar{a}\beta_2} + \bar{c}Ay - \bar{b}ABs \\
&= \frac{-\bar{\delta}\tilde{b}/\bar{a}}{\tilde{b} + \beta_1\bar{b} + \beta_2\bar{a}} \left(\frac{\bar{a}}{\tilde{b}}\, \tilde{s} - Ay \right) + \left(\frac{-\bar{\delta}\beta_1}{\tilde{b} + \bar{b}\beta_1 + \bar{a}\beta_2} - \bar{b} \right) ABs - \left(\frac{\bar{\delta}(\beta_2 + \tilde{b}/\bar{a})}{\tilde{b} + \bar{b}\beta_1 + \bar{a}\beta_2} - \bar{c} \right) Ay \\
&= \frac{-\tilde{b}\sqrt{\bar{\delta}}}{\bar{a}\sqrt{\eta}(\tilde{b} + \bar{b}\beta_1 + \bar{a}\beta_2)}\hat{w} - \frac{\bar{b}\tilde{b}/\bar{a} + \bar{c}\beta_1 + \bar{b}\beta_2}{\tilde{b} + \bar{b}\beta_1 + \bar{a}\beta_2} v_2 \triangleq \sqrt{\frac{\bar{c}}{\bar{a}}}(\alpha\hat{w} + \beta v_2),
\end{aligned}
$$

using identities

$$
\begin{aligned}
&\bar{\delta}\beta_1 + \bar{b}(\tilde{b} + \bar{b}\beta_1 + \bar{a}\beta_2) = (\bar{b}\tilde{b}/\bar{a} + \bar{c}\beta_1 + \bar{b}\beta_2)\bar{a}, \\
&-\bar{\delta}(\beta_2 + \tilde{b}/\bar{a}) + \bar{c}(\tilde{b} + \bar{b}\beta_1 + \bar{a}\beta_2) = (\bar{b}\tilde{b}/\bar{a} + \bar{c}\beta_1 + \bar{b}\beta_2)\bar{b}.
\end{aligned}
$$

Thus we have

$$
\begin{aligned}
\alpha^2 + \beta^2 &= \frac{\bar{\delta}\tilde{b}^2/\eta + [\bar{b}(\bar{a}\beta_2 + \tilde{b}) + \bar{a}\bar{c}\beta_1]^2}{\bar{a}\bar{c}(\bar{a}\beta_2 + \tilde{b} + \bar{b}\beta_1)^2} \\
&= \frac{\bar{\delta}\tilde{b}^2/\eta + \bar{b}^2(\bar{a}\beta_2 + \tilde{b})^2 + 2\bar{a}\bar{b}\bar{c}\beta_1(\bar{a}\beta_2 + \tilde{b}) + \bar{a}^2\bar{c}^2\beta_1^2}{\bar{a}\bar{c}(\bar{a}\beta_2 + \tilde{b})^2 + 2\bar{a}\bar{b}\bar{c}\beta_1(\bar{a}\beta_2 + \tilde{b}) + \bar{a}\bar{b}^2\bar{c}\beta_1^2} \\
&= 1 - \bar{\delta}[(\bar{a}\beta_2 + \tilde{b})^2 - \bar{a}\bar{c}\beta_1^2 - \tilde{b}^2/\eta]/[\bar{a}\bar{c}(\bar{a}\beta_2 + \tilde{b} + \bar{b}\beta_1)^2] \leqslant 1
\end{aligned}
$$

by (59) and $\bar{\delta} \geqslant 0$. □

Corollary 4.3. *Consider the shifted variable metric method* (49) *satisfying Assumption* 4.2 *with $\underline{\mu}$ sufficiently small and suppose that the line-search method fulfils* (1) *and* (2). *Let the objective function satisfy Assumption* 4.1. *If $\mu^2 \leqslant \zeta\hat{a}/a$ or $\mu = \underline{\mu}$, then* (3) *holds.*

Proof. Choosing $\beta_1 = \beta_2 = 0$ in (58), (59) gives $\eta \geqslant 1$ and it suffices to use Theorem 4.3 with $\eta = 1$. □

This approach cannot be used for method (48), which uses $\beta_2 = 0$ and $\beta_1^2 = \tilde{b}/(\bar{c}\varrho)$ by (47). Then condition (59) is $\tilde{b} - \tilde{b}/\eta \geqslant \bar{a}/\varrho$, which cannot be satisfied in general. Fortunately, similar assertion as Lemma 4.2 holds. Denote again $\tilde{H}_+ = \zeta I + A_+$.

Lemma 4.6. *Let $\bar{\delta} \neq 0$. Consider the shifted variable metric method* (48) *in the form*

$$
U_+ = U - pq^{\mathrm{T}}U, \quad p = \tilde{s} - (\vartheta/\varrho)ABs, \quad q = (y - \vartheta Bs)/p^{\mathrm{T}}y, \tag{60}
$$

with $\vartheta^2 \leqslant \varrho\tilde{b}/\bar{c}$ and $\vartheta\bar{b} \leqslant 0$. Then

$$
\det \tilde{H}_+ / \det H \leqslant (\zeta\hat{a} + \varrho\tilde{b})p^{\mathrm{T}}Bp/\tilde{b}^2. \tag{61}
$$

Proof. Update (60) can be written $A_+ = A - Aqp^{\mathrm{T}} - pq^{\mathrm{T}}A + q^{\mathrm{T}}Aqpp^{\mathrm{T}}$, or

$$\tilde{H}_+ = H^{1/2}\left(I + \frac{B^{1/2}(q^{\mathrm{T}}Aqp - Aq)(q^{\mathrm{T}}Aqp - Aq)^{\mathrm{T}}B^{1/2} - B^{1/2}Aqq^{\mathrm{T}}AB^{1/2}}{q^{\mathrm{T}}Aq}\right)H^{1/2},$$

where $q^{\mathrm{T}}Aq > 0$ by $\bar{a} - 2\vartheta\bar{b} + \vartheta^2\bar{c} = (\vartheta\bar{c} - \bar{b})^2 + \bar{\delta} > 0$. Since $\det(I + (u-v)(u-v)^{\mathrm{T}} - vv^{\mathrm{T}}) = |u|^2 + (1 - u^{\mathrm{T}}v)^2 - |u|^2|v|^2$ (see the proof of Lemma 4.2), we obtain

$$\det \tilde{H}_+/\det H = q^{\mathrm{T}}Aqp^{\mathrm{T}}Bp + (1 - p^{\mathrm{T}}BAq)^2 - p^{\mathrm{T}}Bpq^{\mathrm{T}}ABAq.$$

Observing that $q^{\mathrm{T}}ABAq = q^{\mathrm{T}}Aq - \zeta q^{\mathrm{T}}q + \zeta^2 q^{\mathrm{T}}Bq$ and $1 - p^{\mathrm{T}}BAq = 1 - p^{\mathrm{T}}q + \zeta p^{\mathrm{T}}Bq = (\vartheta/p^{\mathrm{T}}y)p^{\mathrm{T}}Bs + \zeta p^{\mathrm{T}}Bq$, we find by the Schwarz inequality and (60)

$$\begin{aligned}\det \tilde{H}_+/\det H &= p^{\mathrm{T}}Bp[\zeta q^{\mathrm{T}}q - \zeta^2 q^{\mathrm{T}}Bq] + [p^{\mathrm{T}}B((\vartheta/p^{\mathrm{T}}y)s + \zeta q)]^2 \\ &\leqslant p^{\mathrm{T}}Bp[\zeta q^{\mathrm{T}}q - \zeta^2 q^{\mathrm{T}}Bq + ((\vartheta/p^{\mathrm{T}}y)s + \zeta q)^{\mathrm{T}}B((\vartheta/p^{\mathrm{T}}y)s + \zeta q)] \\ &= [\zeta|y - \vartheta Bs|^2 + \vartheta^2 c + 2\zeta\vartheta s^{\mathrm{T}}B(y - \vartheta Bs)]p^{\mathrm{T}}Bp/(p^{\mathrm{T}}y)^2 \\ &= (\zeta\hat{a} + \vartheta^2 c - \zeta\vartheta^2|Bs|^2)\frac{p^{\mathrm{T}}Bp}{(p^{\mathrm{T}}y)^2} = (\zeta\hat{a} + \vartheta^2\bar{c})\,\frac{p^{\mathrm{T}}Bp}{(\tilde{b} - \vartheta\bar{b}/\varrho)^2} \\ &\leqslant (\zeta\hat{a} + \vartheta^2\bar{c})p^{\mathrm{T}}Bp/\tilde{b}^2 \leqslant (\zeta\hat{a} + \varrho\tilde{b})p^{\mathrm{T}}Bp/\tilde{b}^2\end{aligned}$$

and by assumptions. □

Lemma 4.7. *Consider the shifted variable metric method* (60), *satisfying* $|\vartheta| \leqslant \tilde{C}$ *for some* $0 < \tilde{C} < \infty$. *Then* $p^{\mathrm{T}}Bp \leqslant 2\tilde{s}^{\mathrm{T}}B\tilde{s} + 2c(\tilde{C}/\underline{\varrho})^2$.

Proof. Observing that $\zeta s^{\mathrm{T}}B^3 s/s^{\mathrm{T}}B^2 s \leqslant \zeta\|B\| \leqslant 1$, we get $s^{\mathrm{T}}BABABs = c - 2\zeta s^{\mathrm{T}}B^2 s + \zeta^2 s^{\mathrm{T}}B^3 s \leqslant c - \zeta s^{\mathrm{T}}B^2 s \leqslant c$ and therefore

$$p^{\mathrm{T}}Bp = |B^{1/2}(\tilde{s} - (\vartheta/\varrho)ABs)|^2 \leqslant 2[\tilde{s}^{\mathrm{T}}B\tilde{s} + (\vartheta/\varrho)^2 c] \leqslant 2\tilde{s}^{\mathrm{T}}B\tilde{s} + 2c(\tilde{C}/\underline{\varrho})^2. \qquad \square$$

Theorem 4.5. *Consider the shifted variable metric method* (48) *satisfying Assumption* 4.2 *with* μ *sufficiently small and suppose that the line-search method fulfils* (1) *and* (2). *Let the objective function satisfy Assumption* 4.1. *If* $\vartheta_k = -\operatorname{sgn}\bar{b}_k \min[\tilde{C}, \sqrt{\varrho_k\tilde{b}_k/\bar{c}_k}]$, $k \geqslant 1$, *for some* $0 < \tilde{C} < \infty$ *and* $\mu^2 \leqslant \zeta\hat{a}/a$ *or* $\mu = \underline{\mu}$, (3) *holds*.

Proof. Using Lemmas 4.6, 4.3, 4.4, 4.5 and 4.7, we can proceed in the similar way as in the proof of Theorem 4.2. □

4.3. Nonconvex case

Modifying the direction vector, we can assure global convergence in the nonconvex case.

Theorem 4.6. *Let the objective function* $f : \mathscr{R}^N \to \mathscr{R}$ *be bounded from below and have bounded second-order derivatives. Consider the line-search method satisfying* (2) *with*

$$d_k = -H_k g_k - \sigma_k |H_k g_k| g_k, \tag{62}$$

where H_k *is symmetric positive definite*, $k \geqslant 1$. *If* $\sigma_k \geqslant \underline{\sigma} > 0$, $k \geqslant 1$, *then* (3) *holds*.

Proof. Assume, for contradiction purposes, that (3) does not hold. Then we can suppose $|g_k| \geqslant \underline{\varepsilon}$ for some $\underline{\varepsilon} > 0$ and $g_k H_k g_k > 0$, $k \geqslant 1$, by positive definiteness of H_k. Omitting index k, we have from (62) by the Schwarz inequality

$$d^{\mathrm{T}} d \leqslant |Hg|^2 + 2\sigma |Hg|^2 |g| + \sigma^2 |Hg|^2 |g|^2 = (1 + \sigma|g|)^2 |Hg|^2$$

and $-g^{\mathrm{T}} d > \sigma |Hg| |g|^2$. Thus $-g^{\mathrm{T}} d/(|g||d|) > \sigma|g|/(1+\sigma|g|) \geqslant \underline{\sigma\varepsilon}/(1+\underline{\sigma\varepsilon})$, since function $\xi/(1+\xi)$ is increasing. Using Theorem 1.1, we have a contradiction. □

We tested choice (62) with $\sigma_k = \underline{\sigma}$, $k \geqslant 1$, using Test 28 from [12], and found that numerical results were very similar for $\underline{\sigma} \leqslant 10^{-6}$.

5. Conclusions

In this contribution, we describe and analyze a family of shifted variable metric methods and prove their global convergence. These methods, originally developed to generate starting matrices for limited-memory methods, are competitive with the best implementations of the standard variable metric methods as demonstrated in Section 2.4.

Furthermore, we present four new limited-memory methods closely related to the shifted variable metric family and prove their global convergence. Our numerical experiments reported in Section 3.3 demonstrate their efficiency in comparison with the known methods for large-scale optimization.

Acknowledgements

We thank the two anonymous referees for careful reading of the first version and for constructive suggestions.

References

[1] M. Al-Baali, Extra-updates criterion for the limited memory BFGS algorithm for large scale nonlinear optimization, J. Complexity 18 (2002) 557–572.

[2] M.C. Biggs, Minimization algorithms making use of nonquadratic properties of the objective function, J. Inst. Math. Appl. 8 (1971) 315–327.

[3] 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.

[4] E.D. Dolan, J.J. Moré, Benchmarking optimization software with performance profiles, Math. Programming 91 (2002) 201–213.

[5] R. Fletcher, Practical Methods of Optimization, Wiley, Chichester, 1987.

[6] J.Ch. Gilbert, J. Nocedal, Global convergence properties of conjugate gradient methods for optimization, SIAM J. Optim. 2 (1992) 21–42.

[7] P.E. Gill, M.W. Leonard, Limited-memory reduced-Hessian methods for large-scale unconstrained optimization, SIAM J. Optim. 14 (2003) 380–401.

[8] D.C. Liu, J. Nocedal, On the limited memory BFGS method for large scale optimization, Math. Programming 45 (1989) 503–528.

[9] L. Lukšan, C. Matonoha, J. Vlček, LSA: algorithms for large-scale optimization, Report V-896, Prague, ICS AS CR, 2004.

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

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

[12] L. Lukšan, J. Vlček, Test problems for unconstrained optimization, Report V-897, Prague, ICS AS CR, 2003.

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

[14] S.S. Oren, D.G. Luenberger, Self scaling variable metric (SSVM) algorithms, Management Sci. 20 (1974) 845–874.

[15] D.F. Shanno, K.J. Phua, Matrix conditioning and nonlinear optimization, Math. Programming 14 (1978) 144–160.

[16] J. Vlček, L. Lukšan, New variable metric methods for unconstrained minimization covering the large-scale case, Report V-876, Prague, ICS AS CR, 2002, (www.cs.cas.cz/∼luksan/reports.html).

[17] J. Vlček, L. Lukšan, Additional properties of shifted variable metric methods, Report V-899, Prague, ICS AS CR, 2004, (http://www.cs.cas.cz/∼luksan/reports.html).