Math. Program., Ser. A (2012) 135:413–436
DOI 10.1007/s10107-011-0479-6

FULL LENGTH PAPER

A limited memory steepest descent method

Roger Fletcher

Received: 2 December 2009 / Accepted: 3 May 2011 / Published online: 21 July 2011

Abstract The possibilities inherent in steepest descent methods have been considerably amplified by the introduction of the Barzilai–Borwein choice of step-size, and other related ideas. These methods have proved to be competitive with conjugate gradient methods for the minimization of large dimension unconstrained minimization problems. This paper suggests a method which is able to take advantage of the availability of a few additional 'long' vectors of storage to achieve a significant improvement in performance, both for quadratic and non-quadratic objective functions. It makes use of certain Ritz values related to the Lanczos process (Lanczos in J Res Nat Bur Stand 45:255–282, 1950). Some underlying theory is provided, and numerical evidence is set out showing that the new method provides a competitive and more simple alternative to the state of the art l-BFGS limited memory method.

Mathematics Subject Classification (2000) 90C06 · 90C26 · 65K05

1 Introduction

This paper considers the problem of finding an unconstrained local minimizer $\mathbf{x}^*$ of a given continuously differentiable function $f(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^n$, where the gradient vector $\mathbf{g}(\mathbf{x})$ of first partial derivatives is available. The study has been motivated by some ongoing work concerning a Sequential Linear Programming (SLP) algorithm for large scale Nonlinear Programming (NLP), in which a suitable algorithm is required for carrying out unconstrained optimization in the null space. The algorithm would need to be effective for both small and large numbers of variables, and would need to be

R. Fletcher (✉)
Department of Mathematics, University of Dundee,
Dundee DD1 4HN, Scotland, UK
e-mail: fletcher@maths.dundee.ac.uk

matrix-free (in the sense of not storing any potentially large reduced Hessian matrices). Currently the obvious Conjugate Gradient (CG) method has been used, but this has not proved to be very suitable. In particular, in an SLP algorithm, once the correct active set is identified, the null space basis usually changes smoothly as the solution is approached. However, the CG algorithm must be restarted for each SLP iteration, and it is not very convenient to make use of information from the previous SLP iteration.

Also CG algorithms can be very slow, and it would be an advantage if a limited memory approach were available to make use of a limited number, $\overline{m}$ say, of additional 'long' vectors. The l-BFGS method [17] is such a method, and has proved to be very successful in the context of unconstrained optimization, but is less convenient for NLP when changes in the dimension or span of the null space basis take place on each SLP iteration.

I have therefore, returned to some thoughts that I had some 20 years ago [8], occasioned by innovative ideas inherent in the Barzilai–Borwein (BB) methods [1]. These are steepest descent methods with a novel choice of the step-length on each iteration, and have proved considerably superior to the largely ineffective classical steepest descent method [3] in which the step-length is determined by making a line search along the steepest descent direction. Other related choices of step-length have also been proposed since that time (for example [6,11,21,22,26]). What little is known about the theoretical properties of such methods is largely confined to the case in which $f(\mathbf{x})$ is a quadratic function whose Hessian matrix, A say, is positive definite. The relevance of the eigenvalues of A to the analysis is pointed out in [8] and [9], and it is suggested that a limited memory approach might be fashioned by using a limited number of eigenvalue estimates, based on certain Krylov sequence properties. However, the idea was not taken any further at the time, and it is a version of that idea that is explored in this paper.

The methods that we consider are *steepest descent methods*, that is the generic expression for a new iterate $\mathbf{x}^{c+1}$ is

$$\mathbf{x}^{c+1} = \mathbf{x}^c - \alpha_c \mathbf{g}^c \tag{1}$$

in which $\mathbf{g}^c$ refers to $\mathbf{g}(\mathbf{x}^c)$, where $\mathbf{x}^c$ is the current iterate, $\alpha_c > 0$ is a step-length whose choice is determined by the method under consideration, and $-\mathbf{g}^c$ is the *steepest descent direction* at $\mathbf{x}^c$. We recall a result (see for example [7]) that steepest descent methods are invariant under an orthogonal transformation of variables.

The theoretical basis of the methods under consideration derives from the minimization of a positive definite quadratic function, and this is set out in Sect. 2. In the quadratic case, the steplength choice is usually the reciprocal of a *Rayleigh quotient* of the Hessian matrix A. For example the Cauchy choice is

$$\alpha_c^{-1} = \mathbf{g}^{cT} A \mathbf{g}^c \Big/ \mathbf{g}^{cT} \mathbf{g}^c \tag{2}$$

and the Barzilai–Borwein choice is

$$\alpha_c^{-1} = \mathbf{g}^{(c-1)T} A \mathbf{g}^{c-1} \Big/ \mathbf{g}^{(c-1)T} \mathbf{g}^{c-1}. \tag{3}$$

(It is noted in passing that Barzilai and Borwein also consider a second steplength formula based on the Rayleigh quotient

$$\alpha_c^{-1} = \mathbf{g}^{(c-1)T} A^2 \mathbf{g}^{c-1} \Big/ \mathbf{g}^{(c-1)T} A \mathbf{g}^{c-1} \tag{4}$$

and we return to consider this in Sect. 7.) In Sect. 2 a new concept of a *sweep method* is described. This method might be considered as a generalization of the BB method (3), insofar as (3) corresponds to the case $\overline{m} = 1$ in the new method. Numerical evidence indicates that a worthwhile improvement over the BB method can be gained by increasing $\overline{m}$. As with the BB method, the basic sweep method is non-monotonic in regard to both the sequences $\{f^c\}$ and $\{||\mathbf{g}^c||\}$. Convergence can be proved using a generalization of the approach given by Raydan [19], and this is set out in the Appendix.

In Sect. 3, again for a quadratic function, modifications of the basic sweep method are suggested that provide a certain monotonicity property in regard to $\{f^c\}$. The motivation is to provide a mechanism for forcing convergence that will carry over to the non-quadratic case. Some numerical evidence indicates that these modifications do not slow down the rate of convergence of the algorithm. Further difficulties arise when the algorithm is generalised to minimize a non-quadratic function, and suggestions for dealing with them are made. In Sect. 5 an implementation of the new algorithm is tested against other common gradient methods, and in particular against the l-BFGS method, with satisfactory results. It is well known that improvements in the performance of CG methods can be obtained for certain types of problem by the use of *preconditioning*. It is likely that similar improvements can be obtained with the sweep methods of this paper. The necessary changes are set out in Sect. 6, and it is shown that the extra storage and housekeeping demands are modest. Conclusions are drawn in Sect. 8, along with suggestions for future work.

Unless otherwise specified in the paper, $||\cdot||$ refers to the L_2 vector norm.

2 The quadratic case

In this section we consider the case in which $f(\mathbf{x})$ is a positive definite quadratic function. For theoretical purposes we may also take $\mathbf{x}^*$ to be the zero vector and so express

$$f = \tfrac{1}{2}\mathbf{x}^T A\mathbf{x} \qquad \text{and} \qquad \mathbf{g} = A\mathbf{x} \tag{5}$$

where A is a symmetric positive definite matrix.

To analyse the convergence of any SD method in this case, we can assume without loss of generality that an orthogonal transformation has been made that transforms A to a diagonal matrix $\Lambda = \text{diag}(\lambda_i)$ of its eigenvalues. Moreover, if there are any eigenvalues of multiplicity $m > 1$, then we can choose the corresponding eigenvectors so that the initial (transformed) gradient vector, $\mathbf{g}^1$ say, has $g_i^1 = 0$ for at least $m-1$ of its corresponding components. It follows from (1) and (5) that gradients recur according to

$$\mathbf{g}^{c+1} = \mathbf{g}^c - \alpha_c A \mathbf{g}^c \tag{6}$$

or component-wise, when $A = \Lambda$, as

$$g_i^{c+1} = (1 - \alpha_c \lambda_i) g_i^c \qquad i = 1, 2, \ldots, n. \tag{7}$$

Clearly, if g_i^c is zero for some i, then this property is preserved by the iteration formula. Also if $\alpha_c = \lambda_i^{-1}$ at any time, then $g_i^{c+1} = 0$ for all subsequent iterations, and such indices have no further effect on the convergence of the algorithm. Thus we can ignore such indices and assume without any loss of generality that Λ has distinct eigenvalues

$$0 < \lambda_1 < \lambda_2 < \cdots < \lambda_n, \tag{8}$$

and that

$$g_i^c \neq 0 \qquad i = 1, 2, \ldots, n \tag{9}$$

on all iterations. A simple consequence of (7) is

A Finite Termination Property for SD: If the eigenvalues of A are known, and if $\alpha_{c+i-1} = \lambda_i^{-1}$ is chosen for $i = 1, 2, \ldots, n$ then $\mathbf{g}^{c+n} = \mathbf{0}$.

In order to fashion a limited memory method, we shall assume on any iteration that in addition to $\mathbf{x}^c$ and $\mathbf{g}^c$, the most recent m back values,

$$G = \left[\mathbf{g}^{c-m}, \ldots, \mathbf{g}^{c-2}, \mathbf{g}^{c-1}\right] \tag{10}$$

say, are also available in wrap-around storage, where $m \leq \overline{m}$ is limited to an upper bound $\overline{m}$ on the number of such vectors that can be stored. When n is large, it is assumed that $\overline{m} \ll n$. The resulting method can thus be implemented (also in the non-quadratic case below) with $\overline{m} + 2$ long vectors.

An important property, possessed by all steepest descent methods (1), is that

$$\mathbf{x}^c - \mathbf{x}^{c-m} \in \text{span}\left\{\mathbf{g}^{c-m}, A\mathbf{g}^{c-m}, A^2\mathbf{g}^{c-m}, \ldots, A^{m-1}\mathbf{g}^{c-m}\right\}. \tag{11}$$

That is to say, the displacement of the current iterate $\mathbf{x}^c$ from any back value $\mathbf{x}^{c-m}$, lies in the span of the so-called *Krylov sequence* initiated from $\mathbf{g}^{c-m}$. It follows that

$$\mathbf{g}^c - \mathbf{g}^{c-m} \in \text{span}\left\{A\mathbf{g}^{c-m}, A^2\mathbf{g}^{c-m}, \ldots, A^m\mathbf{g}^{c-m}\right\}. \tag{12}$$

A remarkable property of this Krylov sequence is that it provides m distinct estimates (so-called *Ritz values*) of the eigenvalues of A, which are contained in the spectrum of A in a certain optimal sense. The theory of this is very extensive, relating to the CG and Lanczos methods, see for example Golub and Van Loan [12].

The Lanczos iterative process [15], applied to the matrix A, starting from $\mathbf{q}_1 = \mathbf{g}^{c-m}/||\mathbf{g}^{c-m}||$, generates orthonormal basis vectors $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_k$ for the

Krylov sequence (11) of any dimension k. Because the columns of G are in this Krylov sequence, we may express $G = QR$ where Q is the matrix with columns $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_m$, such that $Q^T Q = I$, and R is upper triangular and nonsingular, assuming that the columns of G are linearly independent. The Ritz values on which the new method is based are the eigenvalues of the matrix

$$T = Q^T A Q \tag{13}$$

which, as we shall see, is tridiagonal. Under the conditions (8) and (9), the Lanczos process would terminate after exactly n iterations, and so for $m \leq n$, these Ritz values exist, and if $m = n$ they are identical to the eigenvalues of A. Moreover, if $m = 1$, then $Q = \mathbf{q}_1 = \mathbf{g}^{c-m}/||\mathbf{g}^{c-m}||$ and we see from (13) that there is a single Ritz value, namely the Rayleigh quotient (3) on which the BB method is based.

In practice, computation of the Ritz values from the Lanczos process requires the matrix A to be available, which will not be the case when we generalise to non-quadratic functions. An alternative way of computing the Ritz values is the following. Given m back gradients as in (10), we can rearrange the equations arising from (6) in matrix form as

$$AG = [G\, \mathbf{g}^c]J \tag{14}$$

where J is the $(m+1) \times m$ matrix

$$J = \begin{bmatrix} \alpha_{c-m}^{-1} & & & \\ -\alpha_{c-m}^{-1} & \ddots & & \\ & \ddots & \alpha_{c-2}^{-1} & \\ & & -\alpha_{c-2}^{-1} & \alpha_{c-1}^{-1} \\ & & & -\alpha_{c-1}^{-1} \end{bmatrix}. \tag{15}$$

It follows that

$$G^T A G = G^T \left[G\, \mathbf{g}^c\right] J. \tag{16}$$

If the columns of G are linearly independent, we may substitute $G = QR$ in (14) and rearrange using (13), leading to

$$T = \left[R\, Q^T \mathbf{g}^c\right] J R^{-1}. \tag{17}$$

T is readily seen to be upper Hessenberg, and hence tridiagonal and positive definite, since A is symmetric and positive definite. The eigenvalues of T are readily computed, and are the Ritz values referred to above. We denote them by θ_i, $i = 1, 2, \ldots, m$. Because $T = Q^T A Q$, it follows that the Ritz values are contained in the spectrum of A.

Computation of Ritz values in this way requires the computation of QR factors, for example by the modified Gram-Schmidt method. This requires $\sim m^2 n$ flops. Also it

is required to find storage for the columns of Q which are 'long' vectors. A further alternative is to compute R and $\mathbf{r}$ from the partially extended Choleski factorization

$$G^T \left[G\, \mathbf{g}^c \right] = R^T \left[R\, \mathbf{r} \right]. \tag{18}$$

Substituting this equation into (16) and then using $G = QR$ and (13) leads to

$$T = [R\, \mathbf{r}]\, J R^{-1}. \tag{19}$$

This requires only $\sim \frac{1}{2} m^2 n$ flops and no extra long vectors, and is the method that has been used in preparing this paper. However, there are issues relating to ill-conditioning and round-off error that have to be considered (see Sect. 4).

The Finite Termination Property of SD, referred to above, suggests using step-lengths which are the inverse of estimates of the eigenvalues of A. The computation of Ritz values provides a way of obtaining such estimates, and indeed the exact eigenvalues when $m = n$. Although $m = n$ is impracticable for large dimension unconstrained optimization, it does occur in large scale constrained optimization when there are many active constraints and the null space has small dimension, as often happens. Thus the idea suggests itself of using step lengths which are the inverse of certain of the Ritz values. The BB method is then a special case of this method when $m = 1$. When $m > 1$, the question arises as to which Ritz values to choose. A previous idea that I explored of trying to choose a 'best' Ritz value on each iteration did not perform much better than the BB method. To make use of the Finite Termination Property, all the Ritz values need to be used on successive SD steps.

The type of method that is explored in this paper is therefore, based on this basic idea. The sequence of steepest descent iterations is divided up into groups of m iterations, referred to as *sweeps*. At the start of each sweep we denote the current iterate and gradient by $\mathbf{x}^k$ and $\mathbf{g}^k$, and we assume that there are available m Ritz values, $\theta_{j,k-1}$, $j = 1, 2, \ldots, m$ from a previous sweep. Within each sweep we carry out m steepest descent steps

$$\mathbf{x}^{j+1,k} = \mathbf{x}^{j,k} - \alpha_{j,k} \mathbf{g}^{j,k} \qquad j = 1, 2, \ldots, m \tag{20}$$

starting from $\mathbf{x}^{1,k} = \mathbf{x}^k$, and using step lengths $\alpha_{j,k} = (\theta_{j,k-1})^{-1}$. Then $\mathbf{x}^{k+1} = \mathbf{x}^{m+1,k}$. Finally the gradients $\mathbf{g}^{j,k}$, $j = 1, 2, \ldots, m$ obtained on the sweep are used to calculate Ritz values for the next sweep. If $n \leq m$, the method terminates at the minimizer of $f(x)$ in (5) after two sweeps. There are still some issues to be addressed, namely how to choose step lengths for the first sweep, and how to order the Ritz values within a sweep. We return to these issues below.

In the notation of (7), if θ_c is the current Ritz value being used, then

$$g_i^{c+1} = \left(1 - \frac{\lambda_i}{\theta_c} \right) g_i^c. \tag{21}$$

Because $\theta_c \in (\lambda_1, \lambda_n)$, we see that $|g_1|$ is monotonically decreasing, but at a potentially slow rate when θ_c is close to λ_n. If however, θ_c is close to λ_1, then $|g_1|$ is

Table 1 Benefits of increasing m in the sweep method

m	$\#sw$	$\#g$
1	235	236
2	111	220
3	73	213
4	48	185
5	31	143
6	24	129
7	23	139
8	18	119

considerably reduced, but

$$\left|g_n^{c+1}\right| = \left|1 - \frac{\lambda_n}{\theta_c}\right| \left|g_n^c\right| \tag{22}$$

increases by a factor close to the condition number of A. Hence, as for the BB method, the sequence of gradient norms $\{||\mathbf{g}^k||\}$ is usually non-monotonic, as also is the sequence of function values $\{f^k\}$. Thus the convergence of the scheme is an issue to be addressed.

It is important to establish that the sweep method improves on the BB method ($m = 1$) as the number of back vectors m is increased. In practice this has always been observed to be the case. A typical example is that shown in Table 1 based on minimizing a quadratic function of 20 variables, with $\lambda_1 = 1$ and the other eigenvalues in geometric progression with ratio $\sqrt{2}$. The gradient components are all initialised to 1, and the calculations are terminated when a value of $||\mathbf{g}^c|| \leq 10^{-6}||\mathbf{g}^1||$ has been achieved. Initialisation and ordering of the Ritz values is carried out as described in the next section. It can be seen that a smooth and worthwhile improvement in the number of sweeps ($\#sw$) and more importantly the number of gradient calls ($\#g$) is obtained as m is increased. However, the method shows some indication of running out of steam at around $m = 7$, which is perhaps a little disappointing, given that for $m = 20$ we might expect to solve the problem with $\#g \approx 40$. A possible reason for this is ill-conditioning in the computation of R as discussed below.

The theoretical properties of the sweep method are also of some interest. In fact, Raydan [19] has proved convergence of the BB method for strictly convex quadratic functions, and it is shown in the Appendix that a similar line of argument can be used to establish convergence of the sweep method. As regards the rate of convergence, Barzilai and Borwein [1] prove that the order of convergence of the BB method ($m = 1$) is superlinear when $n = 2$, and Dai and Fletcher [4] give reasons to believe that this is also true for $n = 3$ but not for $n \geq 4$. An experiment to assess the situation for larger values of m is shown in Fig. 1. In this example, $n = 10$, $\lambda_1 = 1$, and the eigenvalues are in geometric progression with ratio 2. The initial components of the gradient are 1's and m initial Ritz values are specified, which are equally spaced in the interior of the spectrum. The traces show a significant improvement as m increases (note that exact termination would be expected when $m = 10$). Note also the extremely small

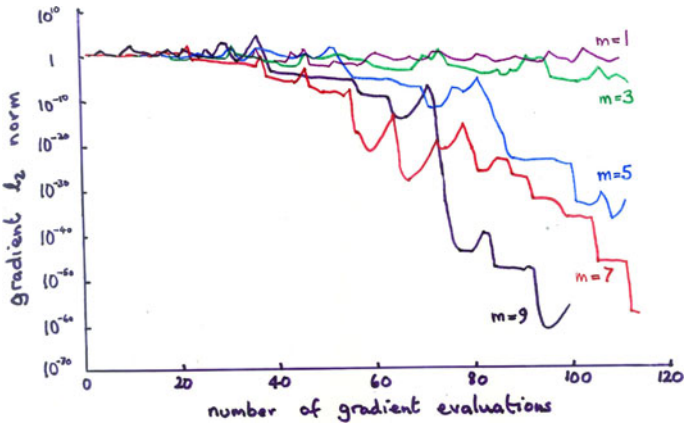

Fig. 1 Convergence of $||\mathbf{g}||$ for $m = 1,\ 3,\ 5,\ 7,\ 9$

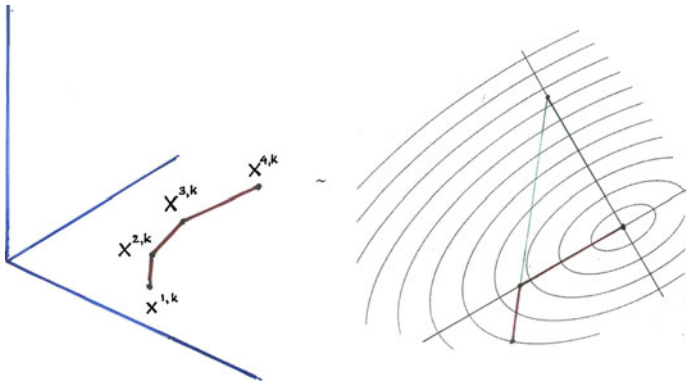

Fig. 2 Visualising a sweep method

values of the gradient norm ($\sim 10^{-60}$ or so) that are achieved as m increases. This can only be observed because the quadratic function in (5) contains no linear terms. Given that the calculations are carried out to a relative precision of 10^{-16} or so, we see that the later iterations are also very effective at reducing rounding errors introduced earlier in the calculation. The negative curvature trend of the traces for $n = 5, 7$ and 9 is suggestive of R-superlinear convergence in exact arithmetic, and we might conjecture that this is so if say $2m \gtrsim n$. However, the technical difficulty of proving such a result would be considerable.

The sweep may be viewed as a *piecewise curvilinear trajectory*, somewhat similar to following the SD trajectory $\dot{\mathbf{x}} = -\mathbf{g}(\mathbf{x})$ with an ODE solver, as illustrated on the left hand diagram of Fig. 2. The right hand diagram illustrates the Finite Termination Property when $n = 2$. The two possible steplengths λ_1^{-1} and λ_2^{-1} are those which yield points on the principal axes of the quadratic. However, monotonic decrease in $f(x)$ is obtained by choosing the smaller steplength first, corresponding to choosing the eigenvalues of A in decreasing order of magnitude.

Table 2 Local behaviour of a sweep method for a non-quadratic function

diag(Λ^*)	78.078	4056.9	16624	70194	495510		
	Moduli of eigencomponents of $\mathbf{g}^{j,1}$						
j	$\|g_1\|$	$\|g_2\|$	$\|g_3\|$	$\|g_4\|$	$\|g_5\|$	f	$\|\mathbf{g}^{j,1}\|$
1	$2.91e{-4}$	$4.73e{-2}$	$6.9e{-2}$	$1.9e{-1}$	$1.3e0$	$2.4e{-6}$	$1.3e0$
2	$2.85e{-4}$	$4.69e{-2}$	$6.7e{-2}$	$1.7e{-1}$	$2.9e{-6}$	$6.0e{-7}$	$1.9e{-1}$
3	$2.84e{-4}$	$4.42e{-2}$	$5.1e{-2}$	$8.9e{-8}$	$1.3e{-5}$	$3.2e{-7}$	$6.8e{-2}$
4	$2.83e{-4}$	$3.34e{-2}$	$1.6e{-7}$	$5.9e{-8}$	$3.9e{-4}$	$1.4e{-7}$	$3.3e{-2}$
5	$2.77e{-4}$	$5.86e{-7}$	$3.2e{-7}$	$1.2e{-7}$	$4.7e{-2}$	$2.8e{-9}$	$4.7e{-2}$
6	$5.52e{-3}$	$1.50e{-2}$	$2.0e{-3}$	$2.3e{-2}$	$3.0e2$	$9.1e{-2}$	$3.0e2$
Replace last step with a line search step							
6	$2.77e{-4}$	$1.06e{-6}$	$3.6e{-8}$	$9.6e{-8}$	$2.5e{-1}$	$6.5e{-8}$	$2.5e{-1}$

The idea of using multiple Ritz values in a sweep method is thought to be new. However, a method of Yuan [26] has the property that it terminates finitely for a 2-dimensional quadratic function. Although the step-length formula is derived in a quite different way to Eqs. (14–19) above, and looks quite different, I think it must be closely related to the case $m = 2$ here.

3 A monotonic sweep method for quadratics

Although the non-monotonic sweep method is effective, the main aim of this paper is to generalize to non-quadratic functions, in which case some attention has to be given to forcing convergence. Moreover, for quadratic programming with box constraints, Dai and Fletcher [5] have given counter examples to show that the BB method can cycle when used in a projection method. So our attention is focussed on deriving a sweep method in which the sequence $\{f^k\}$ decreases *monotonically*, whilst allowing $f^{j,k} > f^{j-1,k}$ within a sweep. First we consider minimizing a positive definite quadratic function.

Our key proposal is to select the Ritz values in decreasing order of size, during a sweep. If $m = n$ we know that the method terminates, in which case selecting the Ritz values in this order ensures that both f and $||\mathbf{g}||$ are monotonically decreased (see [8]). In general, this ordering provides step-lengths $\alpha_{j,k}$ that increase in size as the sweep progresses. This gives the best chance that early steps in the sweep reduce $f^{j,k}$ monotonically. However, if a step fails to improve on the value f^k at the start of sweep, then an interpolation is made to take the Cauchy step in that direction, and the sweep is terminated.

We also terminate a sweep if $f^{j,k}$ improves on f^k, but $||\mathbf{g}^{j,k}|| \geq ||\mathbf{g}^{j-1,k}||$. The reasoning is that higher index components of $\mathbf{g}$ are now growing and the next step in the sweep is likely to fail. Thus we hope to save a possibly unproductive step by this means. An illustration of this is given in Table 2 of Sect. 4.

The algorithm might be summarized as follows

A Ritz Sweep Algorithm

```
Initialize x^c and the stack of Ritz values
For k = 1, 2, ...,
   Denote f^k = f^c
   While the stack is not empty
      Take a Ritz value θ off the stack
      Set α_c = θ^{-1}
      Set x^{c+1} = x^c − α_c g^c
      If f^{c+1} ≥ f^k then
         Reset α_c = g^{cT} g^c / (g^{cT} A g^c)
         Reset x^{c+1} = x^c − α_c g^c and clear the stack
      Else
         If ||g^{c+1}|| ≥ ||g^c|| then clear the stack End
      End
      Set c = c + 1
   End
   Compute up to m̄ new Ritz values
   Place on the stack in increasing order
End
```

A consequence of terminating a sweep early is that fewer than $\overline{m}$ back values of the gradient are available from that sweep. However, we can usually add back gradients from a previous sweep. Thus where possible we always use $\overline{m}$ back gradients when computing Ritz values. When $k \leq \overline{m}$ there are not enough back gradients. We allow the user to initialize from 1 up to $\overline{m}$ Ritz values for the first iteration. If only one value is provided, the first and second sweeps have only one (BB) step. For the third sweep there are two back gradients, and if these provide acceptable steps, the fourth sweep has four back gradients, and so on.

The main question at issue is whether imposing monotonicity in this way reduces the effectiveness of the sweep method in the quadratic case. Omitting step lengths derived from small Ritz values might be expected to cause slow convergence of small index components of the gradient to zero. In practice we have not noticed any significant loss of effectiveness. This is supported by the following example with $n = 1{,}000$, $\lambda_1 = 1$, $\lambda_n = 1{,}000$, and eigenvalues in geometric progression. One Ritz value (1,001/2) is provided initially. Figure 3 shows traces for $\overline{m} = 2$ through 6, which mostly improve as $\overline{m}$ is increased. Note that $\sqrt{f}$ is plotted against the number of gradient evaluations (not sweeps). (Since $f^* = 0$, we may regard $2\sqrt{f}$ as the weighted norm $||\mathbf{g}||_{A^{-1}}$ of the gradient.) Figure 4 shows the corresponding behaviour of $||\mathbf{g}||$ in the case $\overline{m} = 6$, comparing the monotonic algorithm above to the non-monotonic sweep method of Sect. 2. We may observe that the monotonic method is in no way inferior to the non-monotonic method. Also, although $||\mathbf{g}||$ is non-monotonic in both cases, the amplitude of the non-monotonicity in $||\mathbf{g}||$ is significantly less when f decreases monotonically.

Another issue that must be addressed is possible ill-conditioning or even singularity in the matrix R. In exact arithmetic, if the problem satisfies (8) and (9), then

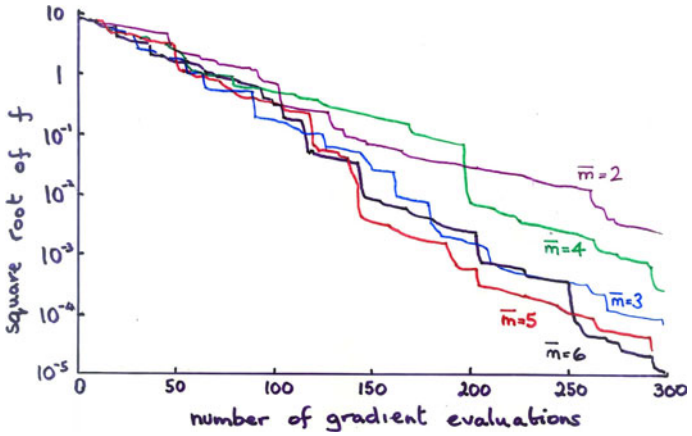

Fig. 3 Sweep-wise monotonic convergence of f for $\overline{m} = 2, \ldots, 6$

the Lanczos process started from any $\mathbf{g}^k$ always takes n steps to terminate, and R is nonsingular for any $m \leq n$. It may be however, that A has multiple eigenvalues, and/or some eigencomponents of the gradient are zero. In this case the Lanczos process terminates in fewer steps, and may not be able to supply $\overline{m}$ Ritz values.

For inexact arithmetic, the back gradients can approach linear dependence as m is increased. Directly computing $G^T G$ doubles the condition number and hence magnifies the effect of round-off error, for example $G^T G$ may become numerically indefinite. Even if QR factors are computed by modified Gram-Schmidt, R can become increasingly ill-conditioned, or even numerically singular, and I have not found that it generates better performance overall. Use of Householder QR would be the most stable method, but is excluded on storage considerations. In any event, for non-quadratic problems, the effect of the non-quadratic terms is much more substantial than that of round-off. In practice it is necessary to monitor the conditioning of R and be prepared to use fewer back values in the computation of R and hence T. For this reason there may be a practical limit to the size of $\overline{m}$, beyond which numerical issues restrict the effectiveness of the scheme. Values of $\overline{m} = 5$, 6 or 7 as in Table 1, seem to be typical.

4 A monotonic sweep method for non-quadratic functions

When the objective function $f(\mathbf{x})$ is non-quadratic, the mechanism for proving convergence set out in the Appendix is no longer valid, and some form of monotonicity is likely to be needed to drive the gradients to zero. Various researchers have recognised this fact. Raydan [20] modifies the BB method in the manner of Grippo et al. [14], requiring sufficient improvement in f^k over the maximum of the back values f^{k-j}, $j = 1, 2, \ldots, M$. Raydan chooses $M = 10$. If sufficient improvement is not obtained, an Armijo line search is carried out. Raydan (private communication) indicates that he occasionally might use smaller values (e.g. $M = 2$ or 3) in case of difficulty, but points out the papers of Varadhan and Gilbert [24] and Vargas et al. [25], where values of $M = 20$ or 50 are used. It is likely that the best choice of M is very

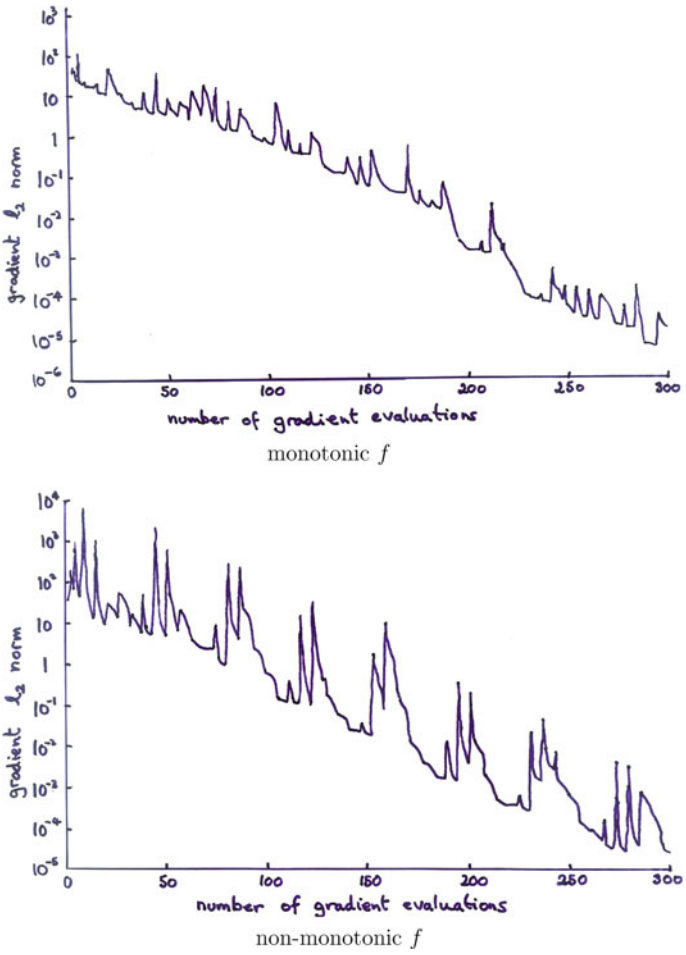

Fig. 4 Convergence characteristics of $||\mathbf{g}||$ for $\overline{m} = 6$

problem dependent. Dai and Fletcher [5] suggest a method that is different to Grippo et al. [14], and possibly more suited to BB methods, for maintaining monotonicity on a subsequence. Dai and Yuan [6] investigate the long term behaviour of various monotone methods, looking for a certain clustering property in the gradients that is possessed by the BB method. Serafini et al. [22] are able to obtain monotonicity by switching between the two BB methods and the Cauchy step, according to certain criteria. Our approach has been to require monotonic decrease of the sequence $\{f^k\}$ in the outer iteration sequence, whilst allowing limited increases in $f^{k,l}$ within a sweep, as in the algorithm above.

The benefit of retaining monotonicity is illustrated by the following non-quadratic example. It is an instance of the Trigonometric Test Problem [10] with $n = 5$. A solution $\mathbf{x}^*$ is designated as a vector of random numbers, uniformly distributed in $\pm\pi$, and the starting point $\mathbf{x}^1$ is obtained by perturbing components of $\mathbf{x}^*$ by uniformly

distributed random numbers in $\pm 10^{-5}$. The eigensolution of the Hessian is calculated at $\mathbf{x}^*$. We see in Table 2 that the condition number $\kappa = 6346$ is quite large, but not unduly so. This amount of information would be expected to provide rapid local convergence from this $\mathbf{x}^1$ with a Newton method, and we would like to see something similar for a sweep method. We use the (orthonormal) eigenvectors to transform the Hessian at $\mathbf{x}^*$ to a diagonal matrix, and the n eigenvalues diag(Λ^*), suitably ordered, are supplied as the initial Ritz values θ_i for an iteration of the sweep method. Under these conditions, for a quadratic problem, exact and monotonic termination would occur in one sweep.

The actual behaviour of the eigencomponents of the gradient (transformed gradients) on the sweep is shown in Table 2. The first step uses the largest Ritz value λ_5, and drives $|g_5|$ almost to zero, as would be expected from a quadratic model. However, there is a small error of $2.9e-6$ due to non-quadratic effects. Similarly on step 2, λ_4 is used and $|g_4|$ is almost zeroed, and so on. We also see a lesser reduction in gradients to the left of the one being almost zeroed. However, gradients to the right start to increase, and the size of the increase is approximately determined by Eq. (22). In particular the increases in $|g_5|$ start to become significant. On the last step, $|g_5|$ has increased to $3.0e2$, resulting in $||\mathbf{g}^2||$ being almost a factor of 200 greater than $||\mathbf{g}^1||$. Moreover, even $|g_1|$ has increased, showing that non-quadratic effects have become dominant. Thus the last non-monotonic step is a disaster as regards providing local convergence. On the other hand, replacing the last step of the sweep by a line search step maintains the good overall progress of the sweep. In fact, because $||\mathbf{g}||$ increases from $j = 4$ to $j = 5$, this is taken as an indication that the next step is likely to be unsatisfactory. Thus in the algorithm above, the sweep would terminate with the $j = 5$ iterate.

There are several issues which must be addressed when generalising to the non-quadratic case, in particular

- the matrix T is necessarily upper Hessenberg, but not usually tridiagonal,
- the matrix A is no longer available to compute the Cauchy step in the algorithm, and
- there are various effects related to the existence of non-positive curvature.

None of these issues admits a single obvious solution, and various possibilities might be followed up.

In regard to the matrix T, it is required to compute some values in a way that reduces to the Ritz values in the quadratic case. Since m is likely to be small, finding all the eigenvalues of T is still an option, but then there is the possibility of complex eigenvalues to be considered. Byrd et al. [2] address similar issues relating to limited memory methods. In practice I have observed that, due to ill-conditioning in R, the product with R^{-1} in (19) seriously magnifies non-quadratic terms in the upper triangle of T. Therefore, my approach has been to construct a symmetric tridiagonal matrix, $\widetilde{T}$ say, essentially by replacing the strict upper triangle of T by the transpose of the strict lower triangle. Then eigenvalues of $\widetilde{T}$ are used to compute Ritz-like values for the next sweep. However, in the case $m = 1$, the Ritz value arising from (19) is

$$\theta_c = \alpha_{c-1}^{-1} \left(\mathbf{g}^{c-1} - \mathbf{g}^c\right)^T \mathbf{g}^{c-1} \Big/ \mathbf{g}^{(c-1)T} \mathbf{g}^{c-1}, \tag{23}$$

which is the general form of the Barzilai–Borwein formula (3) for minimizing a non-quadratic function.

Another issue is that the matrix A is no longer available to compute a Cauchy step. The obvious alternative is to enter some sort of line search when $f^{c+1} \geq f^k$ in the algorithm. One might choose either a combination of Armijo search and interpolation, such as Raydan [20] uses, or a slightly more elaborate search aiming to satisfy Wolfe-Powell conditions (essentially that described in Section 2.6 of Fletcher [7]). I have chosen the latter as it ensures that the resulting $\widetilde{T}$ has at least one positive eigenvalue.

Effects related to non-positive curvature also show up in that the matrix $\widetilde{T}$ may have some negative eigenvalues. Various possibilities suggest themselves. One is to discard the oldest back gradient and recompute a smaller matrix $\widetilde{T}$. Another is simply to put all the Ritz-like values on the stack as before. Then one can either terminate the sweep when a negative Ritz value is found, or carry out a line search before terminating the sweep. Some limited experiments have suggested that there can be a significant difference, and I have preferred the last option.

Finally there is the issue of how best to compute R. Although larger values of m may be possible when using modified Gram Schmidt QR factors, the effect of non-monotonicity, or non-quadratic effects, or negative eigenvalues of $\widetilde{T}$, is that fewer than m of these Ritz values may actually be usable. There are also complications in storing Q, if the provision of extra long vectors is to be avoided. Since computing R as the Choleski factor of $G^T G$ is approximately twice as fast, this is the approach that I have taken.

5 Numerical experience

A Fortran 77 subroutine has been written which implements the ideas of previous sections. It is referred to as `lmsd` (limited memory steepest descent). It is compared with (my) implementations of other standard first derivative methods, on some standard non-quadratic test problems. A more detailed comparison of the limited memory methods lmsd and l-BFGS is also made, including some more difficult CUTEr test problems [13]. The tests are run on a COMPAQ Evo N800v laptop (clock speed 1.3 GHz) under Linux, with optimized code from the Intel F90 compiler.

Other codes tested are the BFGS method, the Polak-Ribiere and Fletcher-Reeves CG methods (see for example [7]), the l-BFGS method [17] and Raydan's non-monotonic BB method [20]. The CG methods use a fairly accurate line search with a two sided test on the slope ($\sigma = 0.1$ as in [7]). The lmsd, BFGS and l-BFGS methods all use the same Wolfe-Powell line search, with $\sigma = 0.9$ in a one sided test on the slope. Test problems include different instances of the Trigonometric Test Problem [10], the Convex2 problem [20], the Chained Rosenbrock Problem [23] starting from $\mathbf{x}^1 = \mathbf{0}$, the Laplace2 problem [9] and various CUTEr test problems [13]. All codes use the same termination condition, namely $||\mathbf{g}^c|| \leq \tau ||\mathbf{g}^1||$. For Tables 3, 4 and 5, $\tau = 10^{-6}$, and for Table 6, $\tau = 10^{-5}$.

For the lmsd runs, only one Ritz value was supplied initially. This was $\theta_1 = 10^5$ for the Trigonometric problems, 10^2 for the Chained Rosenbrock and CUTEr problems, and 1 for the other problems. No attempt was made to optimize over these choices.

Table 3 Trigonometric test problem

	$n = 50$			$n = 100$		
	$\#sw/ls$	$\#f$	$\#g$	$\#sw/ls$	$\#f$	$\#g$
lmsd 2	258	658	473	433	1,115	786
lmsd 3	199	714	583	356	1,281	1,042
lmsd 4	86	304	*265*	152	531	*473*
lmsd 5	111	429	388	141	558	494
lmsd 6	102	512	451	152	679	611
CG PR	558	1,515	1,074	795	2307	1,622
CG FR	319	941	678	628	1,888	1,337
BB-Raydan	1,285	1,780	1,286	2,182	3,142	2,183
l-BFGS 3	731	816	759	1,109	1,265	1,166
l-BFGS 5	719	776	732	1,005	1,113	1,033
BFGS	122	145	130	219	236	224

Table 4 Convex2 test problem

	$n = 10^3$			$n = 10^5$		
	$\#sw/ls$	$\#f$	$\#g$	$\#sw/ls$	$\#f$	$\#g$
lmsd 2	107	271	213	126	326	250
lmsd 3	64	217	185	74	259	214
lmsd 4	39	165	146	50	218	190
lmsd 5	26	126	*114*	39	200	*177*
lmsd 6	29	164	148	34	204	182
CG PR	118	202	194	254	463	402
CG FR	108	221	215	287	387	375
BB-Raydan	172	212	173	260	330	261
l-BFGS 3	132	138	134	210	218	213
l-BFGS 5	117	122	119	232	238	234
BFGS	124	129	126			

Eigenvalues of $\widetilde{T}$ are calculated by the (standard) QL algorithm with implicit shift. For values of $m \sim 6$ and large n, the time taken for this is entirely negligible.

The first set of tests is shown in Table 3 through Table 6. A range of values of $\overline{m}$ is tested for the lmsd method. There is also a similar parameter $\overline{m}$ for the l-BFGS method, and standard values $\overline{m} = 3$ and 5 are tested. For l-BFGS, $2\overline{m} + 4$ long vectors are used in my implementation, as against $\overline{m} + 2$ for lmsd. Note that Raydan's BB method and the CG-FR method require 3 long vectors and CG-PR requires 4. Of course, the BFGS method requires $\frac{1}{2}n^2 + O(n)$ locations, so is not practical for large n. However, where applicable, it shows up as the best method or nearly so in the comparisons, as might be expected, and provides a standard of comparison for the other low storage methods. Excluding BFGS, the run requiring the fewest gradient calls is italicised (Tables 3–7).

Table 5 Chained Rosenbrock test problem

	$n = 50$			$n = 100$		
	$\#sw/ls$	$\#f$	$\#g$	$\#sw/ls$	$\#f$	$\#g$
lmsd 2	1,269	3,319	2,526	1,689	4,360	3,367
lmsd 3	675	2,381	1,981	1,028	3,664	3,029
lmsd 4	853	3,120	2,841	938	3,712	3,301
lmsd 5	558	2,272	2,067	943	4,278	3,772
lmsd 6	608	2,725	2,455	2,490	1,0752	10,265
CG PR	551	1,237	1,000	874	1,842	1,515
CG FR	>9,999			>9,999		
BB-Raydan	>9,999			>9,999		
l-BFGS 3	305	319	308	586	605	589
l-BFGS 5	276	294	*281*	521	541	*525*
BFGS	249	328	280	483	634	540

Table 6 Non-quadratic Laplacian test problem ($n = 10^6$)

	Laplace2 (a)			Laplace2 (b)		
	$\#sw/ls$	$\#f$	$\#g$	$\#sw/ls$	$\#f$	$\#g$
lmsd 2	586	1,553	1,165	503	1,340	1,001
lmsd 3	313	1,167	911	248	873	728
lmsd 4	165	732	633	193	859	746
lmsd 5	128	702	613	102	515	*465*
lmsd 6	111	686	626	100	606	557
CG PR	395	753	640[a]	423	787	675[a]
CG FR	332	597	521[a]	435	660	610[a]
BB-Raydan	1,032	1,395	1,033	1,187	1,620	1,188
l-BFGS 3	495	524	517	521	528	524
l-BFGS 5	395	407	*400*	471	480	474

[a] Failed to achieve a relative improvement of 10^{-5} in $||\mathbf{g}||$

The first observation that emerges is that overall a worthwhile improvement in the performance of the lmsd method is seen as $\overline{m}$ is increased from 2 up to about 5 or 6, beyond which little if any improvement is obtained. This is in line with what was observed in Sect. 3, and the reasons why there is a limit on what can be achieved are probably similar.

Turning to the low storage methods, the lmsd method provides the best method by far for the trigonometric problems. These are quite nonlinear, with a quite large condition number and have no special structure such as sparsity or multiple eigenvalues etc. to take advantage of. This outcome may be an indication that lmsd is likely to perform well on hard but not particularly large problems.

In the Convex2 problem, the objective function is separable, with n distinct eigenvalues in the Hessian at $\mathbf{x}^*$. In terms of gradient counts, lmsd and l-BFGS perform

Table 7 Comparison of limited memory methods

	n	lmsd 5			l-BFGS 3		l-BFGS 5	
		$\#f$	$\#g$	Time	$\#g$	Time	$\#g$	Time
Convex2	10^6	190	*168*	*25.6*	217	81.3	218	104.5
Laplace2a	10^6	702	613	*100.1*	517	199.8	*400*	200.9
Laplace2b	10^6	515	*465*	*75.2*	524	208.4	474	239.4
SPMSRTLS	10^4	330	299	*3.6*	322	4.1	*288*	3.7
NONCVXU2	10^4	9,528	8,381	53.9	*2,161*	*15.1*	2,198	17.0
NONCVXUN	10^4	13,541	11,979	77.8	3,735	*26.3*	*3,732*	28.6
MSQRTALS	1,024	7,491	6,590	61.1	7,283	63.3	*4,310*	*36.3*
MSQRTBLS	1,024	4,751	4,183	39.0	5,267	46.4	*3,140*	*26.8*
QR3DLS	610	81,777	*70,217*	*125.5*	330,316	575.7	176,732	230.0

similarly, with a worthwhile improvement over the CG and BB methods. Note however, the timings below in Table 7, which are considerably in favour of lmsd. This reflects the more simple housekeeping of the lmsd method when the cost of evaluating f and $\mathbf{g}$ is negligible.

The Chained Rosenbrock problem provides a different picture with lmsd showing up badly relative to l-BFGS, and to CG-PR to a lesser extent. It is difficult to provide any very convincing reason for this. My impression is that, due to the way the function is constructed, the gradient path to the solution (as defined by $\dot{\mathbf{x}} = -\mathbf{g}(\mathbf{x})$) has to follow a succession of steep curved valleys, and it is something similar that the lmsd method is doing. Methods which can take large steps may be able to 'jump over' some of the difficulties and hence reach the solution more quickly.

The Laplace2 problem is from a three dimensional p.d.e., with a mildly nonlinear term on the diagonal of the Hessian. An accuracy criterion of $\tau = 10^{-6}$ proved difficult to achieve due to full accuracy in f^* already having been obtained with lower accuracy in $\mathbf{g}$. Thus the termination criterion has been relaxed for this example. The l-BFGS 5 method is a little better on Laplace2 (a), but otherwise there is little to choose, apart from the BB-Raydan method being less successful.

I think these results provide some evidence that the limited extra storage available to the lmsd and l-BFGS methods does in the main lead to improved performance. The next comparison in Table 7 aims to measure the relative performance of these methods. The test set takes in some additional CUTEr test problems which are quite challenging. Mostly these are least square problems derived from a square system of nonlinear equations. Solving these as least square problems might be expected to adversely affect the conditioning of the problem. (If A denotes the Jacobian of the equations, then the Hessian includes a term AA^T). Thus, in order to get reasonably accurate solutions in $\mathbf{x}$, the tolerance on the gradient was decreased to $\tau = 10^{-8}$ for the CUTEr problems. The timings given are in seconds. In the table, for the l-BFGS methods, no function counts are given, as these are marginally greater than the number of gradient counts, except for the QR3DLS problem.

Table 8 Comparison of l-BFGS 5 and 10

	n	l-BFGS 5		l-BFGS 10	
		#g	Time	#g	Time
Convex2	10^6	218	104.5	243	184.4
Laplace2a	10^6	400	200.9	372	282.3
Laplace2b	10^6	474	239.4	383	292.0
SPMSRTLS	10^4	288	3.7	276	4.8
NONCVXU2	10^4	2,198	17.0	2,114	17.3
NONCVXUN	10^4	3,732	28.6	3,564	29.8
MSQRTALS	1,024	4,310	36.3	3,790	33.3
MSQRTBLS	1,024	3,140	26.8	2,742	24.2
QR3DLS	610	176,732	230.0	15,9738	308.9

I have selected $\overline{m} = 5$ as a reasonable compromise as to what can best be achieved with the lmsd approach. This requires 7 long vectors of storage to implement. The choices of $\overline{m} = 3$ and 5 for l-BFGS are usually recommended, and require 10 and 14 long vectors, respectively (at least, in my implementation). Thus the lmsd results are achieved with less storage requirement. These figures are also reflected to some extent in the timings.

The results provide no conclusive outcome either way. The l-BFGS methods mostly do better on the NONCVX and MSQRT problems, but not by more than a factor of about 3. SPMSRTLS is about equal, and the other problems favour lmsd 5, again by a factor of up to 3 or 4. None of the methods fail to solve any of the problems (or those in Table 3 through 6) in reasonable time.

A referee asks that the choice $\overline{m} = 10$ for the l-BFGS method should also be investigated. I give the results for this in Table 8. I also asked Jorge Nocedal for his current assessment of the best choice of $\overline{m}$. He writes that for many years he thought that $\overline{m} = 3$ or 5 was best, but a later study revealed that for some problems, values such as 20, 50 or even higher could be effective, at least in terms of function evaluations. The results of Table 8 support this, showing a modest improvement in gradient counts, albeit at a cost of significant increases in computation time on the higher dimension problems. I do not think that the conclusions regarding lmsd versus l-BFGS are much affected.

6 Preconditioning

It is well established that the performance of conjugate gradient methods on certain types of problem can be improved by the use of preconditioning. Preconditioning has also been successfully used by Molina and Raydan [16] to accelerate the BB method. It seems reasonable to expect therefore, that preconditioning might be used to advantage in the context of the sweep methods described in this paper. This section briefly sets out what would be involved.

The basis of preconditioning is to make a linear transformation of variables

$$\mathbf{y} = L^T \mathbf{x} \tag{24}$$

in which L is nonsingular, with the aim of improving the spectral distribution of some underlying Hessian matrix A. Also L should be easy to calculate, and solves with L should be inexpensive. Gradients and Hessians then transform according to $\mathbf{g} = \mathbf{g}_\mathbf{x} = L\mathbf{g}_\mathbf{y}$ and $A = A_\mathbf{x} = LA_\mathbf{y}L^T$ (see for example [7]). The ideal transformation would be that for which $A_\mathbf{y}$ is the unit matrix, showing that in general we should choose L such that LL^T is an approximation to A. In our case we would then implicitly carry out the steepest descent method $\mathbf{y}^{c+1} = \mathbf{y}^c - \alpha_c \mathbf{g}_\mathbf{y}^c$ in the transformed variables, which maps into

$$\mathbf{x}^{c+1} = \mathbf{x}^c - \alpha_c L^{-T} L^{-1} \mathbf{g}^c \tag{25}$$

in the $\mathbf{x}$ variables. The step lengths α_c are to be the inverses of Ritz values computed from the Choleski factor of $G_\mathbf{y}^T \left[G_\mathbf{y} \, \mathbf{g}_\mathbf{y}^c \right]$, as indicated in (18) and (19). Thus the simplest way to implement the preconditioned sweep method is to store the transformed gradients $\mathbf{g}_\mathbf{y} = L^{-1}\mathbf{g}_\mathbf{x}$. Assuming that $L\mathbf{g}_\mathbf{y} = \mathbf{g}_\mathbf{x}$ can be solved in situ (such as when L is triangular) then no additional storage is needed, other than what is needed to store L. An additional solve with L^T is needed to update $\mathbf{x}^c$ as in (20), and this may require an extra long vector to implement.

Alternatively, a symmetric positive definite matrix B which approximates A^{-1} may be available. Then we have $B = L^{-T}L^{-1}$ and the update formula (25) becomes

$$\mathbf{x}^{c+1} = \mathbf{x}^c - \alpha_c B\mathbf{g}^c, \tag{26}$$

and the Choleski factor of $G^T B[G \, \mathbf{g}^c]$ is used to compute Ritz values.

It will be interesting to evaluate the performance of the preconditioned sweep method on problems for which good preconditioners are available, such as when solving certain types of differential equation.

7 The second Barzilai–Borwein formula

One referee asks whether the second Barzilai–Borwein formula based on the Rayleigh quotient (4) is a special case of a different lmsd scheme. That indeed is the case and the outcome is of some interest. The difference between (3) and (4) is that an extra A appears in the inner products that comprise the numerator and denominator. Now the CG method is related to another method, the MINRES method, in exactly the same way, differing only in the use of a scalar product $\mathbf{x}^T A\mathbf{y}$ in place of $\mathbf{x}^T\mathbf{y}$. A benefit of the MINRES method is that it is applicable to solve systems $A\mathbf{x} = \mathbf{b}$ in which A is symmetric and nonsingular but indefinite. An important problem of this type is the *KKT system*.

$$A = \begin{bmatrix} B & C \\ C^T & O \end{bmatrix}. \tag{27}$$

Now the Ritz values in Sect. 2 are the eigenvalues θ_i of the matrix T in (13). These eigenvalues are in fact the roots of a monic polynomial $\mathcal{P}_m(\lambda)$ that relates the residual

$\mathbf{r}_{m+1}$ in the CG method to the initial residual $\mathbf{r}_1$ through the equation $\mathbf{r}_{m+1} = \mathcal{P}_m(A)\mathbf{r}_1$. Thus the Ritz values play an important role in interpreting the behaviour of the CG method. We might write these eigenvalues $\Theta = \text{diag}(\theta_i)$ as being determined by the eigensystem

$$\left(Q^T A Q\right) X = \left(Q^T Q\right) X\Theta. \tag{28}$$

If we were to include an extra A in the innerproducts we would obtain a generalised eigensystem

$$\left(Q^T A^2 Q\right) X = \left(Q^T A Q\right) X\Theta. \tag{29}$$

The eigenvalues Θ of this system are referred to by Paige et al. [18] as *harmonic Ritz values*. Analogous to the above, they determine polynomials which describe the behaviour of residuals in the MINRES method. In the case $m = 1$ there is just one harmonic Ritz value which is that given by the second BB formula (4).

For $m > 1$, the matrix $T = Q^T A Q$ can be found as in (19), and in a similar way, the matrix $P = Q^T A^2 Q$ can be found from the equation

$$P = R^{-T} J^T \begin{bmatrix} R & \mathbf{r} \\ & \rho \end{bmatrix}^T \begin{bmatrix} R & \mathbf{r} \\ & \rho \end{bmatrix} J R^{-1} \tag{30}$$

where $\begin{bmatrix} R & \mathbf{r} \\ & \rho \end{bmatrix}$ is the Choleski factor of $[G\ \mathbf{g}^c]^T[G\ \mathbf{g}^c]$. We observe that P is a penta-diagonal matrix. The possibility therefore, arises of computing harmonic Ritz values for use in a limited memory scheme. If A is indefinite, then it is possible that T is indefinite, and *reciprocals* of the harmonic Ritz values should be computed from

$$TX = PX\,\text{diag}(\alpha_i). \tag{31}$$

as this generalised eigensystem has the positive definite matrix P on the right hand side.

8 Summary and discussion

The main aim of this project has been to investigate what benefit can be gained in smooth unconstrained minimization from storing a limited number of back values of the gradient vector in a steepest descent method. The approach has been to make use of Ritz values implicit in the Krylov sequence. On the basis of the variety of numerical evidence provided, I feel it is reasonable to conclude that a substantial benefit is available, and that the performance of the resulting method(s) is comparable for large scale systems to what can be obtained from the l-BFGS method. Moreover this is achieved with less extra storage and housekeeping cost.

In applications to non-quadratic problems, it is seen to be important to preserve some sort of monotonicity property in order to be assured of global convergence.

The indications are that this does not interfere with the underlying effectiveness of the unmodified method for a quadratic function.

I see sweep methods as being useful in a number of situations. One is in large scale systems of elliptic p.d.e's, both linear and nonlinear, possibly taking advantage of preconditioning. Another is in projection methods for large scale box constrained optimization, both for quadratic and non-quadratic objective functions. Likewise, in active set methods for general linearly constrained optimization, it is attractive to have the possibility of both termination or rapid convergence for small null spaces, and yet good performance when the null space is large. Finally, for nonlinear programming, when the null space basis changes slowly from one iteration to the next, the Ritz vectors from one iteration are likely to remain beneficial, and provide a simple and convenient way of carrying forward curvature information from one iteration to the next.

It is a little disappointing that there seems to be a limit to the number of back vectors that can be utilised effectively. It is conjectured in Sect. 3 that this may be due to numerical loss of rank in the bundle of back vectors, in which case there may be nothing that can usefully be done. Another explanation might be that adequate coverage of the spectrum of A can be achieved by only a few Ritz values. Fortunately the suggested choice of $\overline{m} = 5$ is a not unreasonable value for the number of extra long vectors that might be available in a large scale application.

Acknowledgments I am very grateful for helpful comments from the two anonymous referees, and from Jorge Nocedal, Marcos Raydan and Valeria Simoncini.

Appendix: a convergence theorem

In this Appendix a theorem is proved that the basic non-monotonic sweep method of Sect. 2 converges when applied to minimize a strictly convex quadratic function. It is assumed that the transformations of Sect. 2 have been carried out, and that (8) and (9) are valid. The theorem follows a similar type of argument to that of Raydan [19]. First we may usefully prove the following lemma.

Lemma *Let* $\{a_k\}$ *be a sequence of positive numbers, let* $\varepsilon > 0$ *be arbitrarily small, and let* $c < 1$ *and* $C \geq 1$ *be positive constants. If*

$$a_k < \varepsilon \;\Rightarrow a_{k+2} \leq C a_{k+1} \tag{32}$$

$$a_k \geq \varepsilon \;\Rightarrow a_{k+2} \leq c a_{k+1} \tag{33}$$

then $\{a_k\}$ *converges to zero.*

Proof Let (32) and (33) hold. If $a_k \geq \varepsilon$ for all k sufficiently large, then (33) leads to a contradiction. Any group of terms for which $a_k \geq \varepsilon$, $a_{k+1} < \varepsilon$ and $a_{k+2} \geq \varepsilon$ is also excluded, since (33) is again contradicted. Hence terms for which $a_k < \varepsilon$ must occur in groups of two or more. Let $a_{k-1} < \varepsilon$, $a_k < \varepsilon$ and $a_{k+1} \geq \varepsilon$. It follows that $a_{k+1} \leq C\varepsilon$. Moreover, if $a_{k+2} \geq \varepsilon$ then it follows that $a_{k+2} \leq C^2\varepsilon$. However, for any subsequent terms the bound is contracted, until the next term with $a_{k+j} < \varepsilon$ is

reached. Thus $a_k \leq C^2\varepsilon$ for all k sufficiently large. Because ε is arbitrarily small, the sequence $\{a_k\}$ converges to zero. □

We now come to the main theorem. In this we use various bounds on how the components of the gradient propagate. These are all simply derived from the equations

$$g_i^{c+1} = \left(1 - \frac{\lambda_i}{\theta_c}\right) g_i^c \qquad i = 1,\ 2,\ \ldots,\ n \tag{34}$$

for a single step, where θ_c is the Ritz value being used, as in (21).

Theorem *The sequence $\{\mathbf{g}^k\}$ generated by the basic $m-$step sweep method ($m \geq 1$) either terminates at, or converges to, the zero vector.*

Proof We need only consider the case that the sequence does not terminate. First we show that $\left\{g_1^k\right\}$ converges to zero. It is a property of the Krylov sequence that Ritz vectors lie in the interval $(\lambda_1,\ \lambda_n)$. It follows for a sweep of m steps that

$$\left|g_1^{k+1}\right| \leq \left(1 - \frac{\lambda_1}{\lambda_n}\right)^m \left|g_1^k\right|, \tag{35}$$

and hence $\left\{g_1^k\right\}$ converges to zero.

Now let $p \in [2, n+1]$ be the largest integer such that the sequences $\left\{g_1^k\right\}, \left\{g_2^k\right\}, \ldots, \left\{g_{p-1}^k\right\}$ all converge to zero. If $p = n+1$, the theorem is proved, so we consider $p \leq n$ and seek to establish a contradiction. Because $\left\{g_p^k\right\}$ does not converge, there exists $\bar{\varepsilon} > 0$ such that for all $\varepsilon \in (0,\ \bar{\varepsilon}]$, there exists an infinite subsequence S such that $\left|g_p^k\right| \geq \varepsilon, k \in S$ and $\left|g_p^k\right| < \varepsilon, k \notin S$. We consider any such value of ε.

Consider the computation of Ritz values for $k \in S$. Because $||\mathbf{g}^k|| \geq \varepsilon$, any accumulation point $\mathbf{q}^\infty$ of the sequence $\mathbf{q}^k = \mathbf{g}^k/||\mathbf{g}^k||, k \in S$, has $q_1^\infty = q_2^\infty = \ldots = q_{p-1}^\infty = 0$, so Ritz values computed from $\mathbf{q}^\infty$ would lie in $[\lambda_p,\ \lambda_n]$. It therefore, follows by continuity of eigenvalues that we can find an iteration number $k_\varepsilon \in S$ for which

$$\theta_{k,l} \in \left(\tfrac{2}{3}\lambda_p,\ \lambda_n\right) \qquad l = 1,\ 2,\ \ldots,\ m \tag{36}$$

for all $k \in S, k \geq k_\varepsilon$. These Ritz values are used on iteration $k+1$, so it follows for such k that

$$\left|g_p^{k+2}\right| \leq c_p^m \left|g_p^{k+1}\right|, \tag{37}$$

where

$$c_p = \max\left(\frac{1}{2},\ 1 - \frac{\lambda_p}{\lambda_n}\right) < 1. \tag{38}$$

Now we consider the entire sequence $\left\{g_p^k\right\}$. Iterations with $k \in S$ have $\left|g_p^k\right| \geq \varepsilon$, and provide the bound (37). Iterations $k \notin S$ have $\left|g_p^k\right| < \varepsilon$, for which we only have the bound

$$\left|g_p^{k+2}\right| \leq C^m \left|g_p^{k+1}\right|, \tag{39}$$

where

$$C = \left|\frac{\lambda_n}{\lambda_1} - 1\right|. \tag{40}$$

If $C < 1$, it follows immediately that the sequence $\left\{g_p^k\right\}$ converges to zero. If $C \geq 1$ we may invoke the above lemma, again showing that the sequence converges. But this contradicts the definition of p. Thus the theorem is proved. □

Remark Using bounds derived from (37), it also follows that the intermediate gradients on each sweep converge to zero.

References

1. Barzilai, J., Borwein, J.M.: Two-point step size gradient methods. IMA J. Numer. Anal. **8**, 141–148 (1988)
2. Byrd, R.H., Nocedal, J., Schnabel, R.B.: Representations of Quasi-Newton matrices and their use in limited memory methods. Math. Progr. **63**, 129–156 (1994)
3. Cauchy, A.: Méthode générale pour la résolution des systèms d'équations simultanées. Comp. Rend. Sci. Paris **25**, 536–538 (1847)
4. Dai, Y.H., Fletcher, R.: On the asymptotic behaviour of some new gradient methods. Math. Progr. **103**, 541–559 (2005)
5. Dai, Y.H., Fletcher, R.: Projected Barzilai–Borwein methods for large-scale box-constrained quadratic programming. Numer. Math. **100**, 21–47 (2005)
6. Dai, Y.H., Yuan, Y.: Analysis of monotone gradient methods. J. Ind. Manag. Optim. **1**, 181–192 (2005)
7. Fletcher, R.: Practical Methods of Optimization. 2nd edn. Wiley, Chichester (1987)
8. Fletcher, R.: Low storage methods for unconstrained optimization. In: Allgower, E.L., Georg, K. (eds.) Computational Solution of Nonlinear Systems of Equations. Lectures in Applied Mathematics (AMS), vol. 26, pp. 165–179 (1990)
9. Fletcher, R.: On the Barzilai–Borwein method. In: Qi, L., Teo, K., Yang, X. (eds.) Optimization and Control with Applications, Series in Applied Optimization, vol. 96. Kluwer, pp. 235–256 (2005)
10. Fletcher, R., Powell, M.J.D.: A rapidly convergent descent method for minimization. Comput. J. **6**, 163–168 (1963)
11. Friedlander, A., Martínez, J.M., Molina, B., Raydan, M.: Gradient method with retards and generalizations. SIAM J. Numer. Anal. **36**, 275–289 (1999)
12. Golub, G.H., Van Loan, C.F.: Matrix Computations. 3rd edn. The Johns Hopkins Press, Baltimore (1996)
13. Gould, N.I.M., Orban, D., Toint, Ph.L.: CUTEr (and SifDec), a constrained and unconstrained testing environment, revisited. ACM Trans. Math. Softw. **29**, 373–394 (2003)
14. Grippo, L., Lampariello, F., Lucidi, S.: A nonmonotone line search technique for Newton's method. SIAM J. Numer. Anal. **23**, 707–716 (1986)
15. Lanczos, C.: An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. J. Res. Nat. Bur. Stand. **45**, 255–282 (1950)

16. Molina, B., Raydan, M.: Preconditioned Barzilai–Borwein method for the numerical solution of partial differential equations. Numer. Algoritm. **13**, 45–60 (1996)
17. Nocedal, J.: Updating quasi-Newton matrices with limited storage. Math. Comput. **35**, 773–782 (1980)
18. Paige, C.C., Parlett, B.N., van der Vorst, H.: Approximate solutions and eigenvalue bounds from Krylov subspaces. Numer. Linear Algebra Appl. **2**, 115–133 (1995)
19. Raydan, M.: On the Barzilai and Borwein choice of steplength for the gradient method. IMA J. Numer. Anal. **13**, 321–326 (1993)
20. Raydan, M.: The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem. SIAM J. Optim. **7**, 26–33 (1997)
21. Raydan, M., Svaiter, B.F.: Relaxed steepest descent and Cauchy-Barzilai–Borwein method. Comput. Optim. Appl. **21**, 155–167 (2002)
22. Serafini, T., Zanghirati, G., Zanni, L.: Gradient projection methods for quadratic programs and applications in support vector machines. Optim. Methods Softw. **20**, 353–378 (2005)
23. Toint, Ph.L.: Some numerical results using a sparse matrix updating formula in unconstrained optimization. Math. Comput. **32**, 839–852 (1978)
24. Varadhan, R., Gilbert, P.D.: BB: an R package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function. J. Stat. Softw. **32**, 1–26 (2009)
25. Vargas, W.E., Azofeifa, D.E., Clark, N.: Retrieved optical properties of thin films on absorbing substrates from transmittance measurements by application of a spectral projected gradient method. Thin Solid Films **425**, 1–8 (2003)
26. Yuan, Y.: A new stepsize for the steepest descent method. J. Comput. Math. **24**, 149–156 (2006)