On the Asymptotic Behaviour of some New Gradient Methods *

Yu-Hong Dai † Roger Fletcher ‡

Numerical Analysis Report NA/212, March 2003

Abstract

The Barzilai-Borwein (BB) gradient method, and some other new gradient methods have shown themselves to be competitive with conjugate gradient methods for solving large dimension nonlinear unconstrained optimization problems. Little is known about the asymptotic behaviour, even when applied to $n-$dimensional quadratic functions, except in the case that $n = 2$. We show in the quadratic case how it is possible to compute this asymptotic behaviour, and observe that as n increases there is a transition from superlinear to linear convergence at some value of $n \geq 4$, depending on the method. By neglecting certain terms in the recurrence relations we define simplified versions of the methods, which are able to predict this transition. The simplified methods also predict that for larger values of n, the eigencomponents of the gradient vectors converge in modulus to a common value, which is a similar to a property observed to hold in the real methods. Some unusual and interesting recurrence relations are analysed in the course of the study.

1 Introduction

The Barzilai-Borwein gradient method [5] and some other new gradient methods for both symmetric positive definite linear systems and unconstrained optimization perform much better than the classical steepest descent method [6] in practical computations. However, little is known about the asymptotic behaviour of such methods, except in

*This work was supported by the EPRSC in UK (no. GR/R87208/01) and the Chinese NSF grant (no. 10171104).

†State Key Laboratory of Scientific and Engineering Computing, Institute of Computational Mathematics and Scientific/Engineering computing, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, PO Box 2719, Beijing 100080, PR China. Email: dyh@lsec.cc.ac.cn

‡Department of Mathematics, University of Dundee, Dundee DD1 4HN, Scotland, UK. Email: fletcher@maths.dundee.ac.uk

the 2−dimensional case. In this paper, we study the asymptotic behaviour in higher dimensions.

We focus on the minimization of a strongly convex quadratic function of n variables

$$f(\mathbf{x}) = \tfrac{1}{2}\mathbf{x}^T A\mathbf{x} + \mathbf{b}^T\mathbf{x}, \tag{1.1}$$

where $A \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix. Suppose in some iterative method that $\mathbf{x}^{(k)}$, $k = 0, 1, 2, \ldots$ is the current point and $\mathbf{g}^{(k)}$ is the gradient of f at $\mathbf{x}^{(k)}$

$$\mathbf{g}^{(k)} = A\mathbf{x}^{(k)} + \mathbf{b}. \tag{1.2}$$

Then a gradient method for solving (1.1) calculates the next point from

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha_k\mathbf{g}^{(k)}, \tag{1.3}$$

where α_k is a stepsize that depends on the method being used.

In the classical steepest descent (SD) method, the stepsize α_k is chosen such that $f(\mathbf{x})$ is minimized along the line $\mathbf{x}^{(k)} - \alpha\mathbf{g}^{(k)}$, yielding

$$\alpha_k^{SD} = \mathbf{g}^{(k)T}\mathbf{g}^{(k)}/\mathbf{g}^{(k)T}A\mathbf{g}^{(k)}. \tag{1.4}$$

Without loss of generality we can take $\mathbf{b} = \mathbf{0}$, in which case use of (1.4) enables one to establish the inequality

$$f(\mathbf{x}^{(k+1)}) \leq \left(\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}\right)^2 f(\mathbf{x}^{(k)}), \tag{1.5}$$

where λ_1 and λ_n are the maximal and minimal eigenvalues respectively of A. Unless the extreme eigenvalues are close, (1.5) suggests the possibility of slow convergence, which indeed is observed in practical computations.

The Barzilai-Borwein (BB) gradient method [5] computes the stepsize α_k according to

$$\alpha_k^{BB} = \mathbf{g}^{(k-1)T}\mathbf{g}^{(k-1)}/\mathbf{g}^{(k-1)T}A\mathbf{g}^{(k-1)}. \tag{1.6}$$

There is also another stepsize choice described by Barzilai and Borwein, that is

$$\alpha_k^{BB'} = \mathbf{g}^{(k-1)T}A\mathbf{g}^{(k-1)}/\mathbf{g}^{(k-1)T}A^2\mathbf{g}^{(k-1)}, \tag{1.7}$$

although practical experience is somewhat in favour of (1.6).

Another class of gradient methods is the cyclic SD stepsize (CSDS) gradient method, as first proposed in [11]. Given some fixed integer $m \geq 2$, the method determines the stepsize by

$$\alpha_k = \begin{cases} \alpha_k^{SD} & \text{if } \operatorname{mod}(k, m) = 0, \\ \alpha_{k-1} & \text{otherwise.} \end{cases} \tag{1.8}$$

The choice of $m = 2$ is of particular interest for in this case we have that $\alpha_k = \alpha_k^{SD}$ for even k and $\alpha_k = \alpha_k^{BB}$ for odd k, so that the method uses the SD stepsize and the BB stepsize alternately. The method is discussed in [7] and [19], and we refer to it as the Alternate Step (AS) method.

Until the development of these new gradient methods, the method of choice for large scale unconstrained optimization has been the conjugate gradient method (possibly preconditioned). For quadratic functions, this method is optimal in a certain sense, so it cannot be expected that the BB and other new gradient methods would be superior. However, for non-quadratic problems, or if significant round-off errors are present, then the new methods become more competitive. Raydan [17] shows how to globalize the BB method using the non-monotone line search technique of [13], and reports a wide range of numerical experience on problems up to 10^4 variables, showing that the resulting algorithm compares reasonably well against the Polak-Ribière and CONMIN techniques. Due to its simplicity and efficiency, the BB method now has received many useful and successful generalizations and applications, see [2, 3, 4, 7, 9, 11, 12, 14, 17, 18, 19], *etc.* Moreover, the numerical experiments in [7] for unconstrained optimization show that the AS method is a promising alternative to the BB method.

Despite the good numerical performance of the BB and CSDS methods, very little is known about their asymptotic behaviour, except in the case $n = 2$. Thus in this paper we are able to extend our knowledge in this area, ultimately with the aim of explaining the unexpectedly good practical performance of the new methods. First we summarise existing results for gradient methods. For the general $n-$dimensional strongly convex quadratic function, we know from Akaike [1] that the convergence rate of the SD method is $Q-$linear. The BB method is also convergent ([16]) and the convergence rate is $R-$linear ([8]). These convergence results can also be extended to a large collection of gradient methods including the CSDS method (see [11, 7]). However, no realistic estimate of the $R-$linear rate of convergence is given in [8] and [7]. For the $2-$dimensional case, $R-$superlinear convergence results are established for both the BB and the AS methods (see [5, 7]). However, for the classical steepest descent method, it is argued in [15] that the $2-$dimensional case may be quite different from the $n-$dimensional case. The question therefore arises as to how special is the $n = 2$ case for the BB and CSDS methods, and it is this that we explore in this paper.

This paper is organized as follows. In Section 2, we make some observations on the progress of the CSDS method. To make long-term observation possible, we choose to express the iteration formula in terms of the logarithm of the absolute value of each gradient eigencomponent. After a further rearrangement of the iteration formula, we are able to view the long term behaviour, and an interesting pattern emerges. In Section 3, we show how some terms in the iteration formula may be neglected, leading to a simplified algorithm which is able to reproduce the long term behaviour to a certain extent. The recurrence relation of the simplified method is interesting and unusual, and we investigate its asymptotic behaviour. In particular we are able to reproduce a transition from superlinear to linear convergence in the gradient vector as n is increased, which is seen in the real method. In Section 4 we repeat these procedures for the BB method, with similar, albeit somewhat more complicated outcomes. Our results apply to both the BB formulae (1.6) and (1.7), or to any method that combines the two formulae. Relations between the real methods and the simplified algorithms are discussed in Section 5.

2 Computing the Asymptotics of the CSDS method

For any gradient method in the form (1.3), it follows from (1.2) that

$$\mathbf{g}^{(k+1)} = (I - \alpha_k A)\mathbf{g}^{(k)}. \tag{2.1}$$

For the purpose of theoretical analysis, we can assume without loss of generality that the coefficient matrix A is diagonal with distinct eigenvalues,

$$A = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n), \quad \text{with } \lambda_1 > \lambda_2 > \cdots > \lambda_n > 0. \tag{2.2}$$

This is because the gradient method is invariant under any orthogonal transformation and we can combine those gradient components if there are any multiple eigenvalues (for example, see [10]). Then for any gradient method, we have by (2.1) and (2.2) that

$$g_i^{(k+1)} = (1 - \alpha_k \lambda_i) g_i^{(k)} \qquad i = 1,\, 2,\, \ldots, n. \tag{2.3}$$

Also we can assume that $g_i^{(k)} \neq 0$ for all k sufficiently large, since if $g_i^{(k)} = 0$ occurs, it follows from (2.3) that the component remains zero on all subsequent iterations, and hence can be disregarded.

For the CSDS method, the SD stepsize (1.4) is used on m successive iterations, so it is convenient to compress the iteration sequence into groups of m iterations. Using (1.4) and the assumptions of the previous paragraph, we can write

$$g_i^{(mk+m)} = \left(1 - \frac{\lambda_i \sum_{j=1}^n (g_j^{(mk)})^2}{\sum_{j=1}^n \lambda_j (g_j^{(mk)})^2}\right)^m g_i^{(mk)}. \qquad k = 0,\, 1,\, 2, \ldots \tag{2.4}$$

For small values of n, this sequence is observed to converge very rapidly to zero, and underflow of the components to zero soon occurs. Therefore this recurrence is not very convenient for observing the asymptotics of the iteration. A start is made by introducing the vector $\mathbf{p}^{(k)}$ defined by

$$p_i^{(k)} = \log |g_i^{(mk)}|, \tag{2.5}$$

and it follows from $g_i^{(mk)} \to 0$ that $p_i^{(k)} \to -\infty$. Then it follows from (2.4) that

$$p_i^{(k+1)} = p_i^{(k)} + m\left(\log |\textstyle\sum_{j\neq i}(\lambda_j - \lambda_i) \exp 2p_j^{(k)}| - \log(\sum_j \lambda_j \exp 2p_j^{(k)})\right). \tag{2.6}$$

Unfortunately this also does not enable the asymptotics to be seen because the exponentials underflow to zero quite soon, and then the calculation of the log term fails. To circumvent this, we first sort the components of $\mathbf{p}^{(k)}$ into decreasing order, giving a vector

$\mathbf{q}^{(k)}$. That is, we define a vector $\boldsymbol{\pi}^k$ to be a permutation of the numbers $\{1, 2, \ldots, n\}$ such that

$$q_i^{(k)} = p_{\pi_i^k}^{(k)} \qquad i = 1, 2, \ldots, n \tag{2.7}$$

and

$$q_1^{(k)} \geq q_2^{(k)} \geq \ldots q_n^{(k)}. \tag{2.8}$$

Now we can divide out the largest exponential factors in the summations, for example (after omitting superscript k) we may rearrange the last term $\log(\sum_j \lambda_j e^{2p_j})$ in (2.6) as

$$\log(e^{2q_1}(\lambda_{\pi_1} + \textstyle\sum_{j\neq\pi_1} \lambda_j e^{2(p_j-q_1)})) = 2q_1 + \log(\lambda_{\pi_1} + \textstyle\sum_{j\neq\pi_1} \lambda_j e^{2(p_j-q_1)}). \tag{2.9}$$

All the exponents $2(p_j - q_1)$ in the resulting expression are nonpositive and there is no difficulty in computing the log term, even if the exponential underflows. Treating the first log term in (2.6) in a similar way, we obtain formulae for recurring the elements of $\mathbf{p}^{(k)}$. The formula for updating $p_{\pi_1}^{(k)}$ (that is $q_1^{(k)}$) is special, and is

$$p_{\pi_1^k}^{(k+1)} = q_1 + 2m(q_2 - q_1) + m \log \left| \frac{\lambda_{\pi_2} - \lambda_{\pi_1} + \sum\limits_{\substack{j\neq\pi_1 \\ j\neq\pi_2}} (\lambda_j - \lambda_{\pi_1}) e^{2(p_j-q_2)}}{\lambda_{\pi_1} + \sum\limits_{j\neq\pi_1} \lambda_j e^{2(p_j-q_1)}} \right| \tag{2.10}$$

after suppressing superscript k on the right hand side. Otherwise the formula is

$$p_i^{(k+1)} = p_i + m \log \left| \frac{\lambda_{\pi_1} - \lambda_i + \sum\limits_{\substack{j\neq\pi_1 \\ j\neq i}} (\lambda_j - \lambda_i) e^{2(p_j-q_2)}}{\lambda_{\pi_1} + \sum\limits_{j\neq\pi_1} \lambda_j e^{2(p_j-q_1)}} \right|, \qquad i \neq \pi_1. \tag{2.11}$$

Equations (2.10) and (2.11) now enable us to compute the long term behaviour of the vectors $\mathbf{p}^{(k)}$ accurately. Similar manipulations enable us to observe the asymptotics of the BB method, as described in Section 4.

In the case that $n = 2$, these recurrence relations enable us to show that the CSDS method exhibits $m-$step $Q-$superlinear convergence for a strongly convex quadratic, assuming that $p_1^{(0)} \neq p_2^{(0)}$, as documented in [7]. It follows from (2.10) and (2.11) that

$$p_1^{(k+1)} - p_2^{(k+1)} = -(2m-1)(p_1^{(k)} - p_2^{(k)}), \tag{2.12}$$

and hence that π_1^k oscillates between 1 and 2 on alternate ($m-$step) iterations. Thus Equations (2.10) and (2.11) may be expressed using $\lambda_1 > \lambda_2$ as

$$q_2^{(k+1)} = q_1^{(k)} + 2m(q_2^{(k)} - q_1^{(k)}) + c_k \qquad \text{and} \qquad q_1^{(k+1)} = q_2^{(k)} + c_k, \tag{2.13}$$

where

$$c_k = m \log \left(\frac{\lambda_1 - \lambda_2}{\lambda_{\pi_1} + \lambda_{\pi_2} e^{2(q_2^{(k)} - q_1^{(k)})}} \right). \tag{2.14}$$

Because $q_1^{(k)} > q_2^{(k)}$, the exponent in the denominator is negative, and hence c_k may be bounded in modulus by a constant, independent of k. It follows immediately using (2.13) that

$$q_1^{(k+1)} - q_1^{(k)} = -(q_1^{(k)} - q_2^{(k)}) + c_k = -(2m-1)^k (q_1^{(0)} - q_2^{(0)}) + c_k, \tag{2.15}$$

from which $\eta_{k+1}/\eta_k \to 0$ as $k \to \infty$, where η_k denotes $\|g^{(mk)}\|_\infty$, which shows that the rate of convergence is $m-$step $Q-$superlinear. We note that the dominant term in (2.15) does not depend on the eigenvalues and hence the condition number of A.

Moreover, we deduce using an identity, (2.13) and (2.15) that

$$\begin{aligned} q_1^{(k)} &= q_1^{(0)} - \sum_{j=0}^{k-1} (q_1^{(j)} - q_1^{(j+1)}) = q_1^{(0)} - \sum_{j=0}^{k-1} (q_1^{(j)} - q_2^{(j)}) + \sum_{j=0}^{k-1} c_j \\ &= q_1^{(0)} - \sum_{j=0}^{k-1} (2m-1)^j (q_1^{(0)} - q_2^{(0)}) + \sum_{j=0}^{k-1} c_j = -\mu(2m-1)^k + O(k) \end{aligned}$$

after summing the geometric series, where $\mu = (q_1^{(0)} - q_2^{(0)})/(2m-2)$. It then follows that $q_1^{(k+1)} = (2m-1)q_1^{(k)} + O(k)$, or $\eta_{k+1} = \eta_k^{2m-1} e^{O(k)}$. In the case that $m = 2$, this is related to the 'almost' $2-$step cubic convergence result $\|g^{(2k+2)}\| = \|g^{(2k)}\|^{3-\epsilon}$ given in [7].

3 A Simplified CSDS method

The recurrence relations (2.10) and (2.11) also enable us to compute the long-term behaviour of the AS and CSDS methods for $n > 2$. When we do this, we see some very interesting behaviour, illustrated in Figure 1. The almost linear behaviour (on a log scale) for $n = 3$ is strongly suggestive of superlinear convergence of the gradient vectors, but this degrades to much slower convergence for $n = 4$ and $n = 5$. Similar behaviour for $m = 3$ is observed in Figure 2 with a transition from superlinear to linear convergence when $n = 6$. We have also observed the progress for larger values of m and see a transition from superlinear to linear convergence at $n = 2m$. It is of particular interest to understand why the transition occurs in this way.

A trace of the iteration sequence indicates that for $n < 2m$, the log terms in (2.10) and (2.11) become negligible as $k \to \infty$, and the asymptotic behaviour is that generated by the recurrence relations

$$p_{\pi_1^k}^{(k+1)} = q_1^{(k)} + 2m(q_2^{(k)} - q_1^{(k)}) \qquad \text{and} \qquad p_i^{(k+1)} = p_i^{(k)} \quad i \neq \pi_1^k. \tag{3.1}$$

We refer to this as the *simple CSDS method*. If the elements of $\mathbf{p}^{(k)}$ are sorted to give $\mathbf{q}^{(k)}$ as in (2.7) and (2.8), then the effect of an iteration is simply to reduce $q_1^{(k)}$ by $2m(q_1^{(k)} - q_2^{(k)})$ and then re-sort the resulting vector to give $\mathbf{q}^{(k+1)}$. Because $m > 1$, it follows that

$$q_1^{(k+1)} = q_2^{(k)}. \tag{3.2}$$

The simple CSDS method is more readily analysed than the 'real' CSDS method, and this analysis gives some insight into the latter. Particularly we are able to see why the transition from superlinear to linear convergence occurs at $n = 2m$. It can be seen in Figures 1 and 2 that the simple method behaves very similarly to the real method when $n < 2m$, and also shows the same slower convergence when $n \geq 2m$, although in a rather more severe manner. We shall see these features arise in the analysis. Notice however that the simple CSDS method makes no reference to A and its eigenvalues, so there are limits as to what can be deduced from its asymptotic behaviour. For example, we note that the simple CSDS method will stall (that is, fail to make any progress) if $q_1^{(k)} = q_2^{(k)}$, which is not true of the real method.

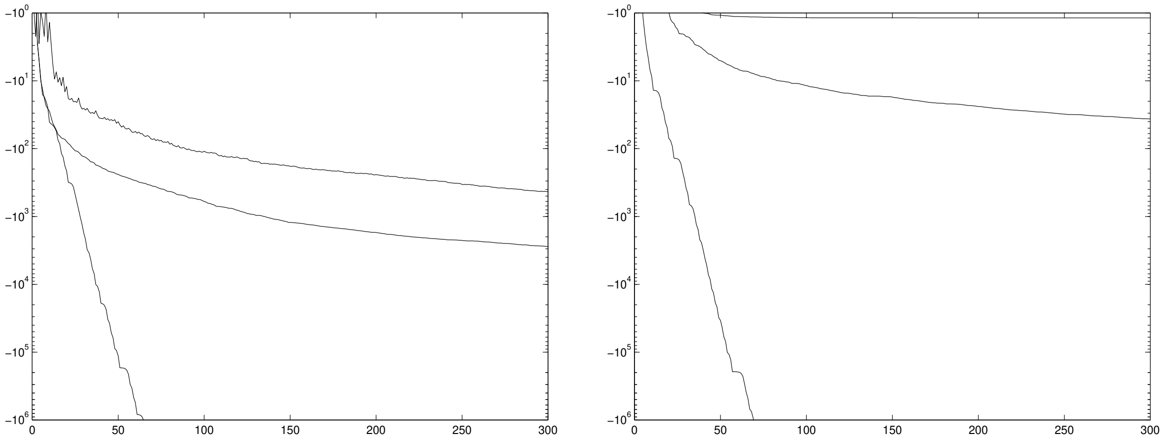

Figure 1: Behaviour of $q_1^{(k)}$ for $n = 3$, 4, 5 in the (a) real, and (b) simple AS methods

To analyse these 'simple' methods, we first introduce two scalars that can be calculated from $\mathbf{q}^{(k)}$. These are

$$\Psi_k = \mathbf{e}^T\mathbf{q}^{(k)} - 2mq_1^{(k)} \qquad \text{and} \qquad \Phi_k = nq_1^{(k)} - \mathbf{e}^T\mathbf{q}^{(k)}, \tag{3.3}$$

where $\mathbf{e}$ is a vector of 1's, and it follows that $\Phi_k = (n-2m)q_1^{(k)} - \Psi_k$. We may also express

$$\Phi_k = \sum_{i=2}^{n}(q_1^{(k)} - q_i^{(k)}) \tag{3.4}$$

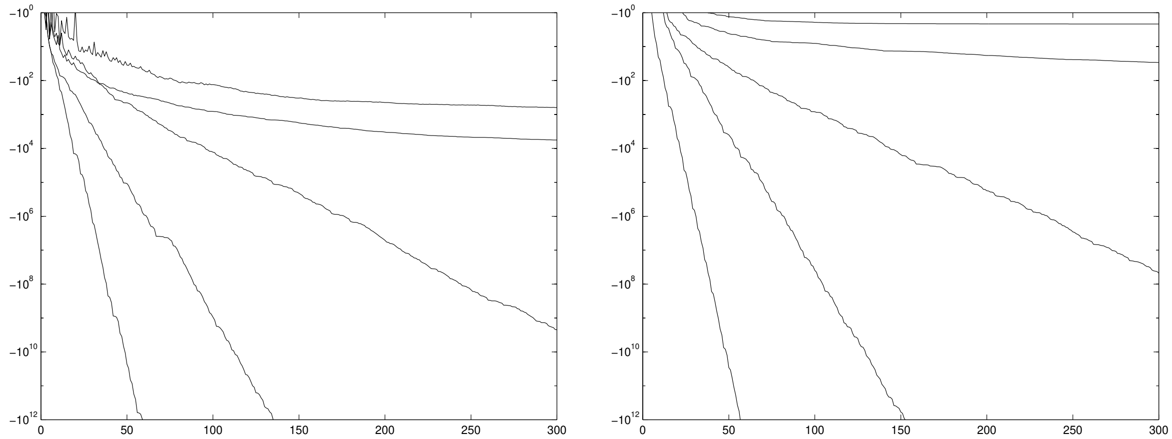

Figure 2: Behaviour of $q_1^{(k)}$ for $n = 3, \ldots, 7$ in the (a) real, and (b) simple CSDS methods ($m = 3$)

from which $\Phi_k > 0$ follows by the ordering of the elements of $\mathbf{q}^{(k)}$. For the simple CSDS method, it follows from the results of the previous paragraph that $\Psi_{k+1} = \Psi_k$, and hence Ψ_k remains constant for all k. It then follows that $\Phi_{k+1} - \Phi_k = (2m - n)(q_1^{(k)} - q_2^{(k)})$. Since $q_1^{(k)} > q_2^{(k)}$, this implies the following property of Φ_k for the simple CSDS method: if $n < 2m$, Φ_k is monotonically increasing; if $n = 2m$, Φ_k stays constant; if $n > 2m$, Φ_k is monotonically decreasing.

Next we investigate the properties of the simple CSDS method, and attempt to relate them to those for the real CSDS method.

Theorem 3.1 *Assume that the simple CSDS method does not stall. Then $q_1^{(k)}$ converges to $-\infty$ if $n \leq 2m$, and fails to converge to $-\infty$ if $n > 2m$.*
Proof For $n > 2m$, because $nq_1^{(k)} > \mathbf{e}^T\mathbf{q}^{(k)}$, it follows that

$$(n - 2m)q_1^{(k)} \geq \mathbf{e}^T\mathbf{q}^{(k)} - 2mq_1^{(k)} = \Psi_k = \Psi_0, \tag{3.5}$$

so $q_1^{(k)}$ is bounded below and does not converge to $-\infty$.

For $n \leq 2m$, we proceed by induction. Clearly the theorem is true for $n = 2$ by setting $c_k = 0$ in (2.15). Now we consider a general value of n ($2 < n \leq 2m$). For any k, there exists a number N_k for which $q_n^{(k)} = q_1^{(k+N_k)}$ on account of the inductive property that $q_j^{(k)} \to -\infty$ for all $j = 1, \ldots, n-1$ when there are $n - 1$ variables. Because $q_1^{(k)}$ is strictly monotonically decreasing, either $q_1^{(k)} \to -\infty$ or $q_1^{(k)} \to q^\infty$. If $q_1^{(k+N_k)} \to q^\infty$ then $q_n^{(k)} \to q^\infty$ and hence $q_j^{(k)} \to q^\infty$ for all $j = 1, \ldots, n$. Thus $\Phi_k \to 0$. But Φ_k is monotonically increasing if $n < 2m$ and constant if $n = 2m$, which is a contradiction. Hence $q_1^{(k)} \to -\infty$. Thus the induction is established, and it follows that $q_1^{(k)} \to -\infty$ whenever $n \leq 2m$. *q.e.d.*

A nice illustration of this result is provided by the following example. For any $n > 2$ and $m \geq 2$, let the number t to be the positive solution of the equation

$$\sum_{i=0}^{n-1} t^i = 2m. \tag{3.6}$$

This solution is unique since the left hand side is monotonic for $t > 0$. Then the values defined by

$$q_i^{(k)} = -\sum_{j=0}^{i+k-1} t^j \tag{3.7}$$

provide a solution to the recurrence relation (3.1). This is readily proved by induction, assuming that the vector $\mathbf{q}^{(0)}$ is initialised accordingly. Assume that the result is true for some $k \geq 0$. Then $q_1^{(k)} - q_2^{(k)} = t^{k+1}$ so $q_1^{(k)}$ is replaced by $q_1^{(k)} - 2mt^{k+1}$, which is $q_1^{(k)} - \sum_{j=k+1}^{k+n} t^j$ from (3.6), or $-\sum_{j=0}^{k+n} t^j$, which agrees with $q_n^{(k+1)}$. Thus the updated $q_1^{(k)}$ becomes $q_n^{(k+1)}$, and hence $q_i^{(k+1)} = q_{i+1}^{(k)}$ for all $i < n$, which again is consistent with (3.7).

The unique solution of (3.6) has $t > 1$ if $n < 2m$, $t = 1$ if $n = 2m$, and $t < 1$ if $n > 2m$. Thus $q_1^{(k)}$ converges to $-\infty$ exponentially if $n < 2m$, linearly if $n = 2m$, and converges to $-1/(1-t)$ if $n > 2m$.

We now move on to analyze the rate at which $\mathbf{q}^{(k)}$ converges to $-\infty$ assuming that $n \leq 2m$. A desirable situation would be to show that there exists a sequence $\{\phi_k\} \to \infty$ such that $q_1^{(k+1)} - q_1^{(k)} \leq -\phi_k$. It then would follow from (2.5) that $\|g^{(k+1)}\|_\infty / \|g^{(k)}\|_\infty \leq \exp(-\phi_k)$. It then follows that the order of convergence is ($m-$step) $Q-$superlinear. However we have not been able to establish this result without making an assumption that in some sense $q_1^{(k)} - q_2^{(k)}$ does not become too close to zero. There are various assumptions that might be used: a simple one is that contained in the following.

Theorem 3.2 *Assume that there exists a constant $\varepsilon > 0$, such that $q_1^{(k)} - q_2^{(k)} \geq \varepsilon(q_1^{(k)} - q_n^{(k)})$ for all k. Then for the simple CSDS method, $q_1^{(k)}$ converges to $-\infty$ exponentially if $n < 2m$, and linearly if $n = 2m$.*

Proof For the simple CSDS method we have by virtue of (3.2), (3.4) and (3.3) that

$$q_1^{(k)} - q_1^{(k+1)} = q_1^{(k)} - q_2^{(k)} \geq \varepsilon(q_1^{(k)} - q_n^{(k)}) \geq \frac{\varepsilon}{n-1}\Phi_k = \frac{\varepsilon}{n-1}((n-2m)q_1^{(k)} - \Psi_k). \tag{3.8}$$

For $n < 2m$, this is the required result.

In the case that $n = 2m$, the right hand side is constant, so convergence is at worst linear. Moreover

$$q_1^{(k)} - q_2^{(k)} \leq \sum_{i=2}^{n}(q_1^{(k)} - q_i^{(k)}) = \Psi_k \tag{3.9}$$

in this case, so convergence is also no better than linear. *q.e.d.*

First we discuss the validity of the assumption made in proving Theorem 3.2. A trace of the ratio $(q_1^{(k)} - q_2^{(k)})/(q_1^{(k)} - q_n^{(k)})$ is shown in Figure 3 for a typical problem with $n = 3$ and $m = 2$. It can be seen that the ratio rarely approaches zero too closely. Hence, even if the assumption is not satisfied for all k, we can conclude that the actual rate of convergence will be very little different from that stated in the theorem.

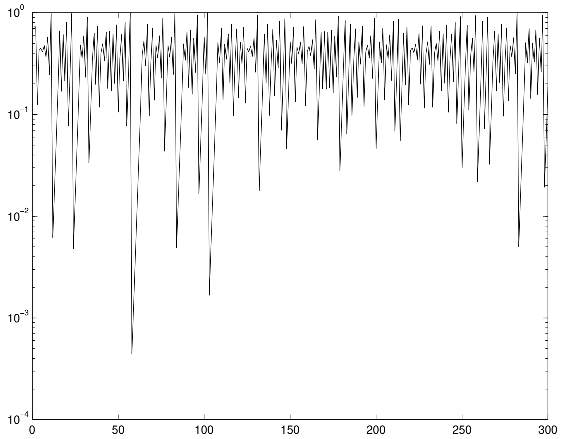

Figure 3: Behaviour of $(q_1^{(k)} - q_2^{(k)})/(q_1^{(k)} - q_3^{(k)})$ for the simple AS method ($n = 3$)

We now discuss to what extent these results have relevance for the real CSDS method, for which $q_1^{(k)} \to -\infty$ for all values of n. The difference between the real and the simple CSDS methods is the presence of the log terms in (2.10) and (2.11). These terms are bounded above by a number that is $O(1)$. In the case that $n < 2m$, we also have $q_1^{(k)} \to -\infty$ for the simple method, so it can be expected that the log terms have negligible effect on the asymptotic rate of convergence. This suggests that the asymptotic behaviour of the real method is also likely to be $m-$step $Q-$superlinear. This is supported by many numerical simulations that we have performed, although we do not yet have a proof of the result. It is possible to prove the result by making the assumption that $\Phi_k \to \infty$, in addition to a separation assumption like that used in Theorem 3.2. However, the only case of which we are aware in which convergence is not $m-$step $Q-$superlinear is when $n = 2$ and $p_1^{(0)} = p_2^{(0)}$. We would therefore hope to prove the result without any assumptions, other than one which excludes this special case.

In the case that $n > 2m$ then it is clear that the simple CSDS method has less to tell us, because $q_1^{(k)}$ converges to a finite limit q_1^{∞}, whereas we know from other work [11, 7] that convergence to $-\infty$ does occur for the real CSDS method. It is a consequence of this that $\Psi_k = \mathbf{e}^T\mathbf{q}^{(k)} - 2mq_1^{(k)} \to -\infty$ contrary to what happens for the simple CSDS method. Clearly these discrepancies are due to the effect of the log terms. However the size of these terms is bounded above by $O(1)$ and we have always observed the terms to vary within a very limited range. Therefore Theorem 3.1 would seem to tell us that although $q_1^{(k)} \to -\infty$, we can expect the differences between the individual terms in $\mathbf{q}^{(k)}$

to stay bounded, and hence for Φ_k to stay bounded. This indeed is what we observe for the real CSDS method.

In the case that $n = 2m$, Theorems 3.1 and 3.2 do predict convergence but at a linear rate. However we have observed that the rate is better than linear in our numerical simulations. It may be due to $\Psi_k \to -\infty$ which we have observed in this case also.

4 The Asymptotics of the BB method

We now carry out a similar analysis of the asymptotic properties of the BB method, with the stepsize choice (1.6), when applied to a strictly convex quadratic function. It is known from the original paper [5] that the convergence is $R-$superlinear for $n = 2$, and a general proof of convergence for all n is given by [16]. To observe the asymptotic behaviour, we first transform to logarithmic coordinates, analogous to (2.5), by defining

$$p_i^{(k)} = \log |g_i^{(k)}|. \tag{4.1}$$

Then, in a similar way to Section 3, we can rearrange the iteration formula (2.3), with $\alpha_k = \alpha_k^{BB}$ in (1.6), giving

$$p_{\pi_1}^{(k+1)} = p_{\pi_1}^{(k)} + 2(q_2 - q_1) + \log \left| \frac{\lambda_{\pi_2} - \lambda_{\pi_1} + \sum\limits_{\substack{j \neq \pi_1 \\ j \neq \pi_2}} (\lambda_j - \lambda_{\pi_1}) e^{2(p_j - q_2)}}{\lambda_{\pi_1} + \sum\limits_{j \neq \pi_1} \lambda_j e^{2(p_j - q_1)}} \right|. \tag{4.2}$$

Here the unsuperscripted entries on the right hand side (including π_1 and π_2) refer to iteration $k-1$, and q_1 and q_2 are the largest and next largest elements of $p^{(k-1)}$. Likewise for $i \neq \pi_1$ the formula is

$$p_i^{(k+1)} = p_i^{(k)} + \log \left| \frac{\lambda_{\pi_1} - \lambda_i + \sum\limits_{\substack{j \neq \pi_1 \\ j \neq i}} (\lambda_j - \lambda_i) e^{2(p_j - q_2)}}{\lambda_{\pi_1} + \sum\limits_{j \neq \pi_1} \lambda_j e^{2(p_j - q_1)}} \right|, \qquad i \neq \pi_1. \tag{4.3}$$

Likewise, if the alternative BB formula in (1.7) is used, a similar rearrangement is possible, that is

$$p_{\pi_1}^{(k+1)} = p_{\pi_1}^{(k)} + 2(q_2 - q_1) + \log \left| \frac{(\lambda_{\pi_2} - \lambda_{\pi_1})\lambda_{\pi_2} + \sum\limits_{\substack{j \neq \pi_1 \\ j \neq \pi_2}} (\lambda_j - \lambda_{\pi_1}) \lambda_j e^{2(p_j - q_2)}}{\lambda_{\pi_1}^2 + \sum\limits_{j \neq \pi_1} \lambda_j^2 e^{2(p_j - q_1)}} \right|, \tag{4.4}$$

and

$$p_i^{(k+1)} = p_i^{(k)} + \log \left| \frac{(\lambda_{\pi_1} - \lambda_i)\lambda_{\pi_1} + \sum\limits_{\substack{j \neq \pi_1 \\ j \neq i}} (\lambda_j - \lambda_i)\lambda_j e^{2(p_j - q_2)}}{\lambda_{\pi_1}^2 + \sum\limits_{j \neq \pi_1} \lambda_j^2 e^{2(p_j - q_1)}} \right|, \qquad i \neq \pi_1. \tag{4.5}$$

These formulae enable us to observe the long term behaviour of the different versions of the BB method.

What we find, arising from (4.2) and (4.3), is illustrated in Figure 4, showing evidence of superlinear convergence when $n = 3$, degrading to linear convergence for $n = 4$ and $n = 5$. If a trace of the iteration sequence is studied, a distinctive pattern emerges, much as for the CSDS method, but rather more complex. This pattern can be reproduced by again ignoring the log terms in (4.2) and (4.3), leading to the recurrence relations

$$p_{\pi_1^{k-1}}^{(k+1)} = p_{\pi_1^{k-1}}^{(k)} + 2(q_2^{(k-1)} - q_1^{(k-1)}) \qquad \text{and} \qquad p_i^{(k+1)} = p_i^{(k)} \quad i \neq \pi_1^{k-1}. \tag{4.6}$$

We refer to this as the *simple BB method*. It can be seen in Figure 4 that the simple method reflects the asymptotic behaviour of the real method to the same extent as for the CSDS method. Again this feature will emerge from the analysis. We note that if the log terms in (4.4) and (4.5) are ignored, then the same simple BB method arises. Thus the consequences of the simple BB method apply to either of the BB formulae (1.6) and (1.7), or to any method using a combination of the two formulae. To ensure that the algorithm is initialised in a consistent way, we shall assume that two vectors $\mathbf{p}^{(0)}$ and $\mathbf{p}^{(1)}$ are given, with $\pi_1^0 = \pi_1^1$ and $q_1^{(1)} = q_1^{(0)} > q_2^{(0)} \geq q_2^{(1)}$.

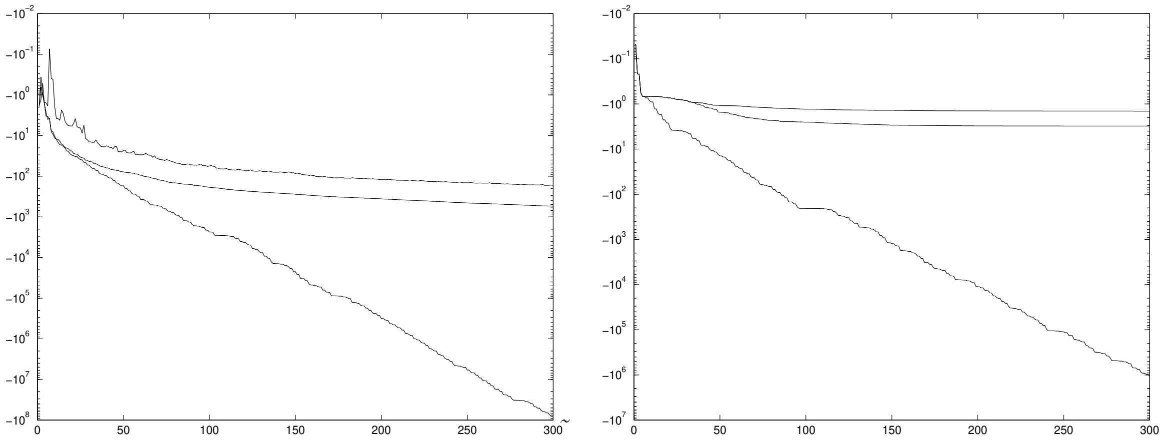

Figure 4: Behaviour of $q_1^{(k)}$ for $n = 3, 4,\ 5$ in the (a) real, and (b) simple BB method

As in the simple CSDS method, only one element of $\mathbf{p}^{(k)}$ changes on each iteration. However the pattern for the simple BB method differs in that the index π_1^k which identifies the largest element of $\mathbf{p}^{(k)}$ does not change on every iteration. In fact we shall show that it stays the same for precisely either two or three successive iterations. Within this group of iterations, we have marked the largest element of $\mathbf{p}^{(k)}$ with a coloured marker: first red, then green and then possibly yellow, when examining print-outs of the simple BB method. We shall use this convention in our analysis. Thus we may denote

$$\mathcal{G} = \{1\} \cup \{k \geq 2 \mid \pi_1^{k-1} \neq \pi_1^{k-2}\} \tag{4.7}$$

as the subsequence of *green iterations*, since it is also appropriate to regard $k = 1$ as a 'green' iteration, due to the method of initialisation.

When $n = 2$, we may define $m_k = p_1^{(k)} - p_2^{(k)}$, and we see from (4.6) that $m_{k+1} = m_k - 2m_{k-1}$. This has the general solution $m_k = A\,2^{k/2}\cos(B + k\arctan\sqrt{7})$, and the constants $A \neq 0$ and B are determined by the initial values of m_0 and m_1. We shall again make use of the function Φ_k defined in (3.3). Since $\Phi_k = |m_k|$ in this case, it readily follows that

$$\frac{\Phi_{k+2}}{\Phi_k} = 2\left|\frac{\cos(C_k + 2\arctan\sqrt{7})}{\cos C_k}\right| \quad \text{and} \quad \frac{\Phi_{k+3}}{\Phi_k} = 2\sqrt{2}\left|\frac{\cos(C_k + 3\arctan\sqrt{7})}{\cos C_k}\right| \tag{4.8}$$

where $C_k = B + k\arctan\sqrt{7}$. It can be shown that the minimum value of these two ratios over all possible values of C_k is equal to 2, and is attained when $C_k = \arctan(1/\sqrt{7}) + i\pi$ for any integer i. Thus there is a subsequence on which $\Phi_k \to \infty$, and hence $q_1^{(k)} \to -\infty$, and convergence is exponential on the subsequence.

At this stage we introduce a more simple notation, local to some $k \in \mathcal{G}$, to aid the presentation of subsequent results. We denote $q_1^{(k-1)} = q_1^{(k)} = a$, $q_2^{(k-1)} = b$, $q_2^{(k)} = c$, and $p_{\pi_1^k}^{(j)} = d,\ e,\ f$ for $j = k+1,\ k+2,\ k+3$ respectively. Thus the values of (q_1, q_2) are (a, b) on iteration $k-1$ and (a, c) on iteration k. On iteration k, a is updated to give $d = a - 2(a - b)$. If $d < c$ then the maximal element changes and $k+1$ is the next red iterate. If $d \geq c$, we regard $k+1$ as a yellow iteration, and update d to get $e = d - 2(a - c)$. We show below that $e < c$, so that in this case $k+2$ is the next red iteration. As for the simple CSDS method, the simple BB method can stall, and this occurs if and only if $a = c$ on some green iteration (this condition also implies $a = b$ on the previous red iteration).

Lemma 4.1 *For all $k \in \mathcal{G}$ it follows that $\pi_1^{k-1} = \pi_1^k$ and either $\pi_1^{k+1} \neq \pi_1^k$ or $\pi_1^{k+2} \neq \pi_1^k$.*
Proof Let $k \in \mathcal{G}$. First we show that $\pi_1^{k-1} = \pi_1^k$. The initialisation ensures that it is true for $k = 1$. If $k > 1$ then $\pi_1^{k-2} \neq \pi_1^{k-1}$, so the π_1^{k-1} element of $\mathbf{p}^{(k-1)}$ (that is a) is not updated on iteration $k-1$. Thus a is unchanged, and therefore is also maximal on iteration k.

If $d < c$ then the maximal element changes and $k+1$ becomes a new red iteration and we have completed a group of two iterations. Moreover, since $e = d - 2(a - c) =$

$c + (d - a) + (c - a)$ and $d < a$, it always follows that $e < c$. Hence $\pi_1^{k+2} \neq \pi_1^k$, so there can be at most three consecutive iterations on which π_1^k remains the same. *q.e.d.*

We now consider the convergence properties of the simple BB method.

Lemma 4.2 *Assume that the simple BB method does not stall. Then for $k \in \mathcal{G}$, Φ_k is monotonic increasing if $n \leq 3$ and monotonic decreasing if $n \geq 4$. Moreover, for $n \geq 4$,*

$$\Phi_{k^+} - \Phi_k \leq (n-4)(q_1^{(k^+)} - q_1^{(k)}), \tag{4.9}$$

where k^+ denotes the successor to k in the subsequence $\mathcal{G}$.

Proof It follows from Lemma 1 that k^+ is either $k+2$ or $k+3$. If $k^+ = k+2$ then $d < c$ and $\Phi_{k^+} - \Phi_k = ((n-1)c - e) - ((n-1)a - c) = (n-3)(c-a) + c - d$. If $k^+ = k+3$, then $d \geq c$ and $\Phi_{k^+} - \Phi_k = ((n-1)c - f) - ((n-1)a - c) = (n-3)(c-a) + d - c$, so both cases can be combined as

$$\Phi_{k^+} - \Phi_k = (n-3)(c-a) + |c-d|. \tag{4.10}$$

Since $a > c$, it follows that Φ_k is monotonic increasing if $n \leq 3$. Alternatively we can express

$$\Phi_{k^+} - \Phi_k = (n-4)(c-a) + \begin{cases} 2(c-b) & \text{if } \ k^+ = k+2 \\ 2(b-a) & \text{if } \ k^+ = k+3 \end{cases}$$

from which the rest of the lemma follows if $n \geq 4$, using $a \geq b \geq c$, $a > c$, and the definitions of a and c. *q.e.d.*

Theorem 4.1 *Assume that the simple BB method does not stall. Then $q_1^{(k)}$ converges to $-\infty$ if $n \leq 3$, and fails to converge to $-\infty$ if $n > 4$.*

Proof For $n > 4$, summing (4.9) over the subsequence of green iterations, it follows that $\Phi_k - \Phi_1 \leq (n-4)(q_1^{(k)} - q_1^{(1)})$, and because $\Phi_k \geq 0$, this provides a finite lower bound on $q_1^{(k)}$.

For $n = 2$, we have already observed that $q_1^{(k)} \to -\infty$. For $n = 3$, we proceed in a similar way to Theorem 3.2. Because the algorithm converges for $n = 2$, it follows for any k, there exists a number N_k for which $q_3^{(k)} = q_1^{(k+N_k)}$ and $k + N_k \in \mathcal{G}$. Because $q_1^{(k)}$ is monotonically decreasing, either $q_1^{(k)} \to -\infty$ or $q_1^{(k)} \to q^\infty$. If $q_1^{(k+N_k)} \to q^\infty$ then $q_3^{(k)} \to q^\infty$ and hence $q_j^{(k)} \to q^\infty$ for all $j = 1, 2, 3$. Thus $\Phi_k \to 0$. But Φ_k is monotonically increasing for $k \in \mathcal{G}$, which is a contradiction. Hence $q_1^{(k)} \to -\infty$ in this case. *q.e.d.*

The only case not covered in Theorem 4.1 is when $n = 4$. In this case, numerical simulations for the simple BB method indicate that $q_1^{(k)} \to q^\infty > -\infty$. Although it is shown in Lemma 4.2 that Φ_k is monotonic decreasing for $k \in \mathcal{G}$, the result is not sufficiently strong to use the argument of Theorem 3.2, and indeed the value of Φ_k can in some cases stay constant from one green iteration to the next.

Simulations of the BB method, both real and simple, appear to show that convergence is better than linear for $n = 3$. As for the simple CSDS method, if we make an assumption that ensures that the stalling case is not approached too closely, we can give a proof of this result. The assumption is that there exists a constant ε, $0 < \varepsilon \ll 1$, such that

$$q_1^{(k)} - q_2^{(k)} \geq \varepsilon(q_1^{(k)} - q_n^{(k)}) \quad \text{for all } k \in \mathcal{G}. \tag{4.11}$$

The logic behind the following proof is based on showing that $\Phi_k \to \infty$ for $k \in \mathcal{G}$, and this is done by first proving a lemma that relates changes in Φ_k to changes in $q_1^{(k)}$ between consecutive green iterations. Now for $n = 3$ it is possible (see (4.10)) that $d = c$ and Φ_k is unchanged. In this case however, a significant increase in Φ_k occurs between the next pair of green iterations, if assumption (4.11) holds. Thus the proof separates into the case when $|c - d|$ is significant and the case when it is close to zero.

Lemma 4.3 *For $k \in \mathcal{G}$, let k^+ and k^{++} denote the next two iteration indices that are in $\mathcal{G}$. If $n = 3$, and assumption (4.11) holds, then*

$$\text{either} \quad \frac{\delta\phi}{\delta q} \geq \varepsilon \qquad \text{or} \qquad \frac{\Delta\phi}{\Delta q} \geq \frac{\varepsilon}{(3 + \varepsilon)}, \tag{4.12}$$

where $\delta\phi = \Phi_{k^+} - \Phi_k$, $\delta q = q_1^{(k)} - q_1^{(k^+)}$, $\Delta\phi = \Phi_{k^{++}} - \Phi_k$ and $\Delta q = q_1^{(k)} - q_1^{(k^{++})}$.

Proof In addition to the simple notation introduced earlier, we also denote $z = q_3^{(k)}$. For the step from k to k^+, it follows as in (4.10) that $\delta\phi = |c - d|$. Also $\delta q = a - c$, so if $|c - d| \geq \varepsilon(a - c)$ it follows that $\delta\phi/\delta q \geq \varepsilon$.

Thus we now consider the case that $|c - d| \leq \varepsilon(a - c)$, and examine the changes $\Delta\phi$ and Δq over two successive green iterations. There are six cases to enumerate, the first three relating to $k^+ = k + 2$ and $c \geq d$, and the second three to $k^+ = k + 3$ and $d \geq c$. Within these categories, three cases relate to the differing relative sizes of z. We denote a^+, b^+ and c^+ to be the values of a, b and c at the start of iteration k^+, and use the fact that the change $\Phi_{k^{++}} - \Phi_{k^+}$ is $|2b^+ - a^+ - c^+|$. In each case we have to determine the sign of the argument, so that the modulus signs can be removed. In all six cases, $c - e = c - (d - 2(a - c)) = 2(a - c) + (c - d)$, $a^+ = c$, and $a - c > 0$.

In the first three cases $0 \leq c - d \leq \varepsilon(a - c)$.

`Case 1:` $d \geq e \geq z$, $b^+ = d$ and $c^+ = e$. Here $2d - c - e = (c - e) - 2(c - d) = 2(a - c) - (c - d) \geq (2 - \varepsilon)(a - c) > 0$. Thus $\Delta\phi = (c - d) + (2d - c - e) = 2(a - c)$, $\Delta q = a - e = (a - c) + (c - e) = 3(a - c) + (c - d)$ and hence $\Delta\phi/\Delta q \geq 2/(3 + \varepsilon)$.

`Case 2:` $d \geq z \geq e$, $b^+ = d$ and $c^+ = z$. Here $2d - c - z = (c - z) - 2(c - d) \geq \varepsilon(c - e) - 2(c - d) = \varepsilon(2(a - c) + (c - d)) - 2(c - d) \geq \varepsilon(c - d)$ using assumption (4.11). Thus $\Delta\phi = (c - d) + (2d - c - z) = (c - z) - (c - d) \geq \varepsilon(c - e) - (c - d) = 2\varepsilon(a - c) - (1 - \varepsilon)(c - d) \geq \varepsilon(a - c)$, $\Delta q = a - z \leq a - e = 3(a - c) + (c - d)$ and hence $\Delta\phi/\Delta q \geq \varepsilon/(3 + \varepsilon)$.

`Case 3:` $z \geq d \geq e$, $b^+ = z$ and $c^+ = z$. Here $2c - d - z = (c - d) + (c - z) \geq 0$. Thus $\Delta\phi = 2(c - d) + (c - z) \geq (c - z)$, $\Delta q = (a - z) \leq (a - e)$ and hence $\Delta\phi/\Delta q \geq \varepsilon$.

In the next three cases $0 \le d-c \le \varepsilon(a-c)$ and $c-f = (c-e)+(e-f) = 2(a-c)+(d-c)$.

`Case 4:` $e \ge f \ge z$, $b^+ = e$ and $c^+ = f$. Here $c+f-2e = (c-e)-(e-f) = 2(a-c)-3(d-c) \ge (2-3\varepsilon)(a-c) > 0$. Thus $\Delta\phi = 2(a-c)-2(d-c)$, $\Delta q = a-f = (a-c)+(c-f) = 3(a-c)+(d-c)$ and hence $\Delta\phi/\Delta q \ge (2-2\varepsilon)/(3+\varepsilon)$.

`Case 5:` $e \ge z \ge f$, $b^+ = e$ and $c^+ = z$. Here $c+z-2e \ge c+f-2e \ge 0$ as for Case 4, and $\Delta q = a-z \ge a-f$ so again $\Delta\phi/\Delta q \ge (2-2\varepsilon)/(3+\varepsilon)$.

`Case 6:` $z \ge e \ge f$, $b^+ = z$ and $c^+ = z$. Here $c+z-2z = c-z \ge 0$. Thus $\Delta\phi = (d-c)+(c-z) \ge c-z$, $\Delta q = a-z = \le a-f$ and hence $\Delta\phi/\Delta q \ge \varepsilon$.

Thus for $\varepsilon > 0$ sufficiently small, $\varepsilon/(3+\varepsilon)$ is a lower bound on $\Delta\phi/\Delta q$. *q.e.d.*

We can now use this lemma to prove a superlinear convergence result for the simple BB method.

Theorem 4.2 *In the simple BB method, if $n = 3$ and assumption (4.11) holds, then $Q-$superlinear convergence of the gradient vectors occurs for $k \in \mathcal{G}$.*

Proof We may define a subsequence of $\mathcal{G}$, initialised by $k = 1$, and subsequently defined by taking the successor of k to be either k^+ or k^{++}, choosing the one which gives the largest ratio in (4.12). Summing the differences, it then follows from (4.12) that if k is an index in the subsequence, then

$$\Phi_k - \Phi_1 \ge \frac{\varepsilon}{(3+\varepsilon)}(q_1^{(1)} - q_1^{(k)}).$$

Because $q_1^{(k)} \to -\infty$, it follows that $\Phi_k \to \infty$ on the subsequence of $\mathcal{G}$, and hence converges monotonically to infinity for all $k \in \mathcal{G}$, by virtue of Theorem 4.1. Then, for all $k \in \mathcal{G}$,

$$q_1^{(k)} - q_1^{(k^+)} = q_1^{(k)} - q_2^{(k)} \ge \varepsilon(q_1^{(1)} - q_3^{(k)}) \ge \tfrac{1}{2}\varepsilon\Phi_k$$

from assumption (4.11) and (3.4). As in the proof of Theorem 3.2, it then follows that $\|\mathbf{g}^{(k^+)}\|_\infty / \|\mathbf{g}^{(k)}\|_\infty \le \exp(-\frac{1}{2}\varepsilon\Phi_k)$, and the result follows because $\Phi_k \to \infty$ for $k \in \mathcal{G}$. *q.e.d.*

The issue of what can be proved for $n = 4$ is more problematic. A nice observation is that provided by the following example. Assume that $\mathbf{q}^{(0)} = -(0,1,1,2)^T$ and $\mathbf{q}^{(1)} = -(0,1,2,3)^T$. Then it is easy to see that $\mathbf{q}^{(2)} = -(1,2,2,3)^T$ and $\mathbf{q}^{(3)} = -(1,2,3,4)^T$. By induction, we deduce for all k that $\mathbf{q}^{(2k)} = -(k,k+1,k+1,k+2)^T$ and $\mathbf{q}^{(2k+1)} = -(k,k+1,k+2,k+3)^T$ Thus it follows that $q_1^{(k)}$ converges to $-\infty$. In this example, the odd numbered iterations are green, and assumption (4.11) is satisfied. Also Φ_k is bounded above, so we only get $Q-$linear convergence of the gradient vectors for $k \in \mathcal{G}$. In fact this example is somewhat unusual, and we actually observe in our numerical experiments that the sequence $\{q_1^{(k)}\}$ generated by the simple BB algorithm with $n = 4$ invariably converges to some $q^\infty > -\infty$.

5 Discussion

We now discuss further the extent to which these results have relevance for the real methods. The situation for the BB method is very similar to that for the CSDS method, namely that the log terms in (4.2) and (4.3) are bounded, so that for $n \leq 3$ it can be expected that the log terms have negligible effect on the asymptotic rate of convergence. This is again supported by many numerical simulations. Thus the simple BB method correctly predicts the observed transition from superlinear to linear convergence of the gradient vectors at $n = 4$.

For $n > 4$ the situation is again similar to that for the CSDS method, because for the simple BB method, $q_1^{(k)}$ converges to a finite limit q_1^{∞}, whereas we know from other work [16] that convergence to $-\infty$ does occur. Clearly these discrepancies are due to the effect of the log terms. Again, the size of these terms is bounded above by $O(1)$ and we have always observed them to vary within a very limited range. Therefore we might extrapolate from Theorem 4.1 that we can expect the differences between the individual terms in $\mathbf{q}^{(k)}$ to stay bounded, even though $q_i^{(k)} \to -\infty$, which is what we observe.

The case that $n = 4$ needs further thought. As indicated above, for the simple BB method, $q_1^{(k)}$ usually but not always converges to a finite limit. For the real method, we know from [16] that $q_1^{\infty} \to -\infty$, and our numerical simulations indicate that convergence is no better than $R-$linear. We do not yet have any very satisfactory theory to support these observations.

The results of this paper are of limited value in predicting practical performance, insofar as the log terms only become negligible well after the problem has been effectively solved to any reasonable degree of precision. This is in contrast to say Newton's method or the BFGS method, where the superlinear convergence of the gradient vectors is very noticeable in practice.

Another point of some interest is that the onset of linear convergence in the CSDS method as n increases can be deferred by increasing the value of m. In practice this strategy is not very effective, certainly not beyond $m = 3$ or $m = 4$. Increasing m increases the extent to which non-monotonic behaviour occurs in the algorithm, and increases the extent to which an eigencomponent of the gradient is eliminated when the Rayleigh quotient α_k^{-1} is close to an eigenvalue. The reason why increasing m is not effective may be that round-off errors limit the extent to which an eigencomponent can actually be eliminated in practice.

References

[1] H. Akaike, *On a successive transformation of probability distribution and its application to the analysis of the optimum gradient method*, Ann. Inst. Statist. Math. Tokyo, 11 (1959), pp. 1-17.

[2] E. G. Birgin, I. Chambouleyron, and J. M. Martínez, *Estimation of the optical constants and the thickness of thin films using unconstrained optimization*, J. Comput. Phys., 151 (1999), pp. 862-880.

[3] E. G. Birgin and Y. G. Evtushenko, *Automatic differentiation and spectral projected gradient methods for optimal control problems*, Optim. Methods Softw., 10 (1998), pp. 125-146.

[4] E. G. Birgin, J. M. Martínez, and M. Raydan, *Nonmonotone spectral projected gradient methods on convex sets*, SIAM J. Optim. 10 (2000), pp. 1196-1211.

[5] J. Barzilai and J. M. Borwein, *Two-point step size gradient methods*, IMA J. Numer. Anal., 8 (1988), pp. 141-148.

[6] A. Cauchy, *Méthode générale pour la résolution des systèms d'équations simultanées,* Comp. Rend. Sci. Paris, 25 (1847), pp. 536-538.

[7] Y. H. Dai, *Alternate Step Gradient Method*, Research report, State Key Laboratory of Scientific and Engineering Computing, Institute of Computational Mathematics and Scientific/Engineering computing, Academy of Mathematics and System Sciences, Chinese Academy of Sciences. 2001.

[8] Y. H. Dai and L.-Z. Liao, *$R-$linear convergence of the Barzilai and Borwein gradient method*, IMA J. Numer. Anal., 26 (2002), pp. 1-10.

[9] R. Fletcher, *Low storage methods for unconstrained optimization,* Lectures in Applied Mathematics (AMS) 26 (1999), pp. 165-179.

[10] R. Fletcher, *On the Barzilai-Borwein Method*, Research report, Department of Mathematics, University of Dundee, 2001.

[11] A. Friedlander, J. M. Martínez, B. Molina, and M. Raydan, *Gradient method with retards and generalizations*, SIAM J. Numer. Anal., 36 (1999), 275-289.

[12] W. Glunt, T. L. Hayden, and M. Raydan, *Molecular conformations from distance matrices*, J. Comput. Chem., 14 (1993), pp. 114-120.

[13] L. Grippo, F. Lampariello, and S. Lucidi, *A nonmonotone line search technique for Newton's method*, SIAM J. Numer. Anal., 23 (1986), pp. 707-716.

[14] W. B. Liu and Y. H. Dai, *Minimization Algorithms Based on Supervisor and Searcher Cooperation*, Journal of Optimization Theory and Applications, 111 (2001), 359-379.

[15] J. Nocedal, A. Sartentaer, C. Zhu, *On the Behavior of the Gradient Norm in the Steepest Descent Method*, Computational Optimization and Applications, 22 (2002), pp. 5-35.

[16] M. Raydan, *On the Barzilai and Borwein choice of steplength for the gradient method*, IMA J. Numer. Anal., 13 (1993), pp. 321-326.

[17] M. Raydan, *The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem*, SIAM J. Optim., 7 (1997), pp. 26-33.

[18] M. Raydan, *Nonmonotone spectral methods for large-scale nonlinear systems*, Report in the International Workshop on "Optimization and Control with Applications", Erice, Italy, July 9-17, 2001.

[19] M. Raydan and B. F. Svaiter, *Relaxed Steepest Descent and Cauchy-Barzilai-Borwein Method*, Computational Optimization and Applications, 21, (2002), pp. 155-167.