Step-Sizes for the Gradient Method*

Ya-xiang Yuan
State Key Laboratory of Scientific/Engineering Computing,
Institute of Computational Mathematics and Scientific/Engineering Computing,
The Academy of Mathematics and Systems Science,
Chinese Academy of Sciences, P.O.Box 2719, Beijing, 100080, P.R.China,
E-mail: yyx@lsec.cc.ac.cn

Abstract

The gradient method searches along the steepest descent direction, the opposite direction of the gradient of the objective function. It can ensure a reduction of the objective function as long as the current iterate point is not a stationary point. Different choices of step-sizes lead to various gradient algorithms. However, the exact line search, which seems to be the best gradient method as it chooses the next iterate by achieving the least function value, turns out to be a very bad method because it often converges very slowly. In the recent years lots of researches have been done on how to choose the step-size for the gradient method, following the amazing result by Barzilai and Borwein(1988), where a specific choice for the step-size is given and proved to ensure superlinear convergence for two dimensional convex quadratic problems. In this paper we review some of the recent advances in this very active and interesting subject, and give new step-sizes for the gradient method.

Keywords: steepest descent, line search, unconstrained optimization, convergence.

1 Introduction

Consider the unconstrained optimization problem:

$$\min_{x\in R^n} f(x), \tag{1.1}$$

where $f(x)$ is a continuously differentiable function in R^n. Let x_k be the current iterate point, and $g_k = g(x_k) = \nabla f(x_k)$ be the gradient vector at x_k. The steepest descent method, which was proposed by Cauchy (1847), defines the next iterate by

$$x_{k+1} = x_k - \alpha_k^* g_k, \tag{1.2}$$

where $\alpha_k^* > 0$ satisfies

$$f(x_k - \alpha_k^* g_k) = \min_{\alpha>0} f(x_k - \alpha g_k). \tag{1.3}$$

*this work is partially supported by Chinese NSF grant 10231060 and by the Knowledge Innovation program of CAS

Cauchy's method is called as the steepest descent method because it can be shown that the steepest descent direction of $f(x)$ at x_k is $-\nabla f(x_k)$. Curry (1944) modified Cauchy's method by replacing α_k^* by

$$\bar{\alpha}_k^* = \min_{\alpha>0}\{\alpha \quad | \quad \nabla f(x_k - \alpha g_k) = 0\}. \tag{1.4}$$

Namely, $\bar{\alpha}_k^*$ is the first stationary point of $f(x)$ along the steepest descent direction $-\nabla f(x_k)$. For strictly convex functions, $\bar{\alpha}_k^*$ and α_k^* are the same.

It is surprising that Cauchy's method is not a good method though it uses the best direction (the direction that descends most) and the best step-size (the step-size that gives the most function reduction). The efficiency of the steepest descent method is first studied by Greenstadt (1967). Assume that $f(x)$ is the following convex quadratic function:

$$f(x) = \frac{1}{2}x^T H x, \tag{1.5}$$

where H is a symmetric positive definite matrix. For the above objective function, it is easy to see the solution $x^* = 0$. Greenstadt (1967) showed that the ratio between the reduction obtained by the Cauchy step and that by the Newton's method is bounded below by

$$\frac{4\mu}{(1+\mu)^2} \tag{1.6}$$

where $\mu = \lambda_1(H)/\lambda_n(H)$ is the ratio between the largest eigenvalue and the smallest eigenvalue of H. Therefore, if the problem is ill-conditioned in the sense that the condition number of the Hessian matrix H is very large, the steepest descent method may converge very slowly. The convergence rate of the steepest descent method was proved by Forsythe (1968):

$$\frac{f(x_{k+1}) - f(x^*)}{f(x_k) - f(x^*)} \leq \left(\frac{\mu - 1}{\mu + 1}\right)^2 < 1. \tag{1.7}$$

We can see that the first inequality in the above relations holds as equality for all k if we choose

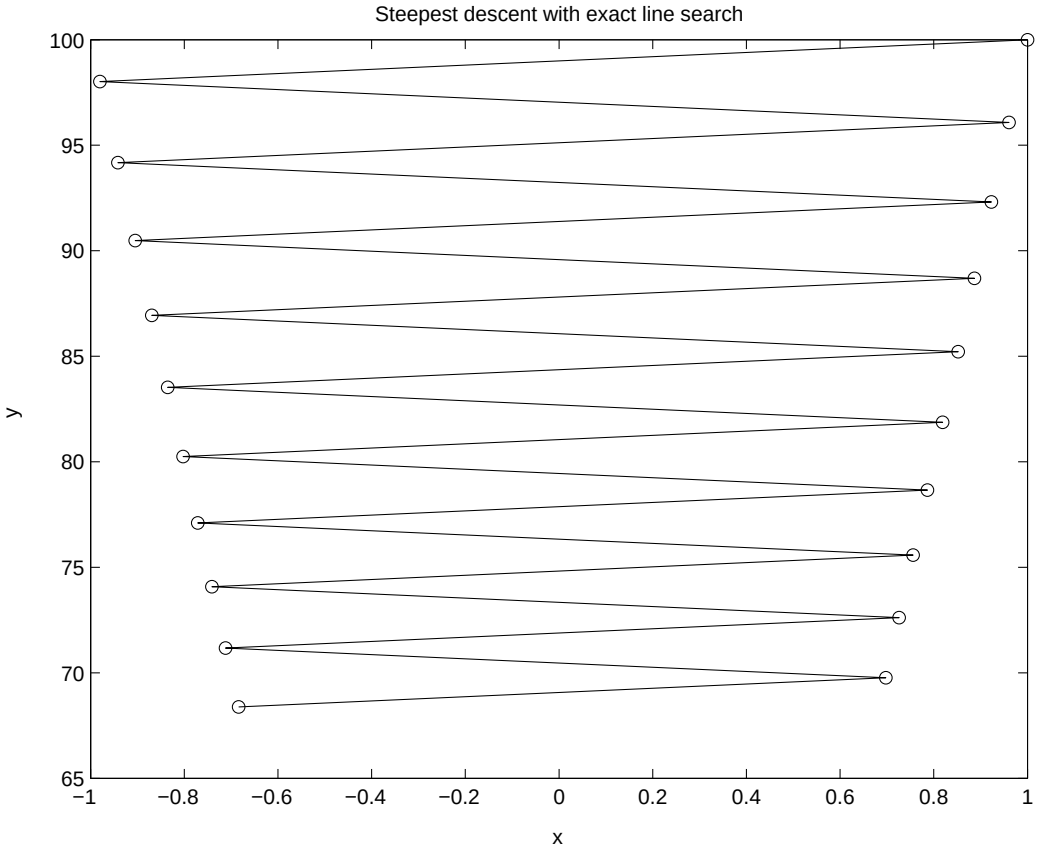
$$H = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}, \tag{1.8}$$

and

$$x_1 = \begin{pmatrix} \frac{1}{\lambda_1} \\ \frac{1}{\lambda_2} \end{pmatrix}, \tag{1.9}$$

and $\lambda_1 > \lambda_2 > 0$. In this example, the iterate points zigzag very slowly, particularly when $\lambda_1 >> \lambda_2$. Though this example is a special problem in 2-dimension, it draws the general picture of the steepest descent method for all n. Akaike (1959) proved that the iterates x_k converge to the solution by asymptotically alternating between two directions - the "cage" of Stiefel (1952), unless the first search direction is an eigenvector of the Hessian H. Furthermore, Akaike (1959) showed that the two asymptotic directions are in a two-dimensional subspace spanned by two eigenvectors of H. Therefore, the steepest descent method converges only linearly and can be very slow if there is a very large ratio between the two eigenvalues whose corresponding eigenvectors span the two dimensional subspace containing the two asymptotic directions. A typical behavior of the steepest descent method is illustrated in the

following picture where 20 iterates are plotted for the objective function $f(x, y) = 100x^2 + y^2$, starting at the initial point (1, 100).



In a practical implementation, instead of exact line search (1.3), we can compute α_k by some line search conditions, such as Goldstein conditions or Wolfe conditions (see Fletcher, 1987). It is easy to show that the steepest descent method with such conditions is always convergent. That is, theoretically the method will not terminate unless a stationary point is found. However, as the exact line search step-size α_k^* normally satisfies such inexact line search conditions, we can see the zigzag phenomenon will also happen.

A surprising result was given by Barzilai and Borwein (1988), which presented formulae for the step-size α_k which lead to superlinear convergence if the objective function is a convex quadratic function of two variables.

The result of Barzilai and Borwein (1988) has triggered off many researches on the gradient method. For example, see Dai (2001), Dai and Fletcher (2003), Dai et al. (2002), Dai and Yuan (2003,2005), Dai and Zhang (2001), Fletcher (2001), Friedlander et al. (1999), Nocedal et al. (2000), Raydan (1993, 1997), Vrahatis et al. (2000), and Yuan (2004). In this paper, we review the recent advances on the general gradient method

$$x_{k+1} = x_k - \alpha_k g_k, \tag{1.10}$$

focusing on the different choices of the step-sizes α_k.

2 The BB Method

The main idea of Barzilai and Borwein's approach is to use the information in the previous interation to decide the step-size in the current iteration. The iteration (1.10) is viewed as

$$x_{k+1} = x_k - D_k g_k, \tag{2.1}$$

where $D_k = \alpha_k I$. In order to force the matrix D_k to have certain quasi-Newton property, it is reasonable to require either

$$\min \|s_{k-1} - D_k y_{k-1}\|_2 \tag{2.2}$$

or

$$\min \|D_k^{-1} s_{k-1} - y_{k-1}\|_2, \tag{2.3}$$

where $s_{k-1} = x_k - x_{k-1}$ and $y_{k-1} = g_k - g_{k-1}$, because in a quasi-Newton method we have that $x_{k+1} = x_k - B_k^{-1} g_k$ and the quasi-Newton matrix B_k satisfies the condition

$$B_k s_{k-1} = y_{k-1}. \tag{2.4}$$

Now, from $D_k = \alpha_k I$ and relations (2.2)-(2.3) we can obtain two step-sizes:

$$\alpha_k = \frac{s_{k-1}^T y_{k-1}}{\|y_{k-1}\|_2^2}, \tag{2.5}$$

and

$$\alpha_k = \frac{\|s_{k-1}\|_2^2}{s_{k-1}^T y_{k-1}} \tag{2.6}$$

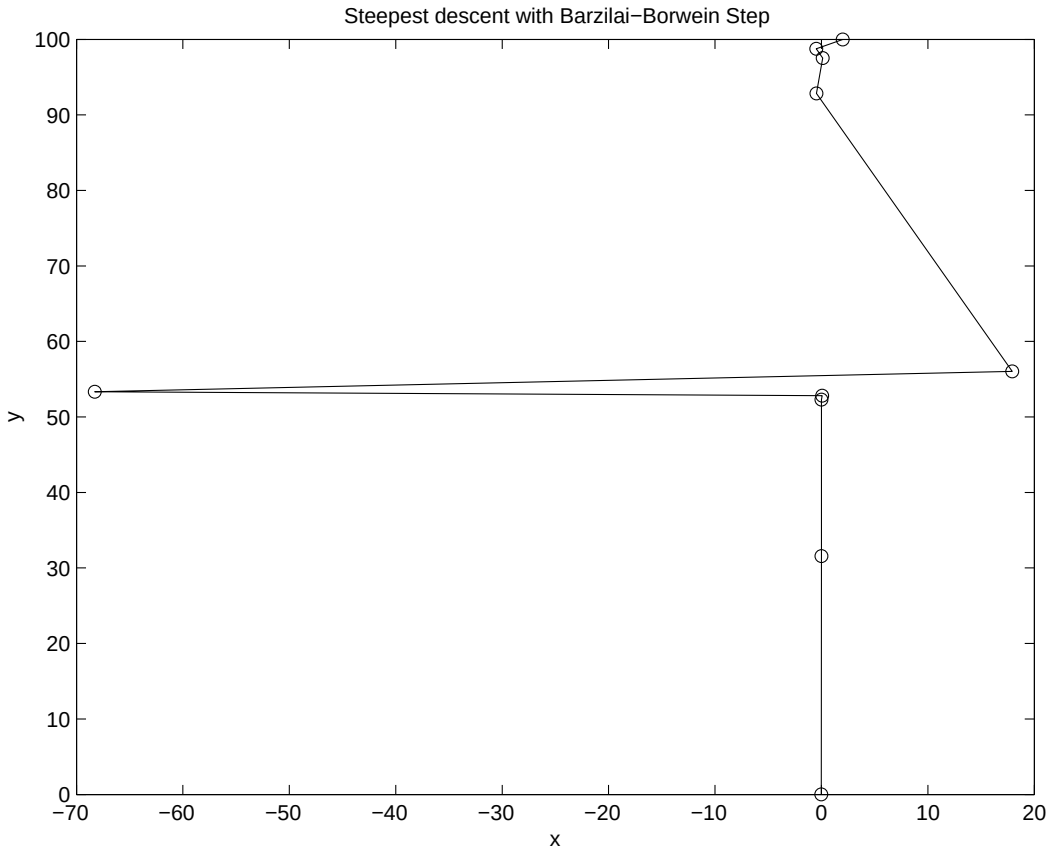
respectively. For convex quadratic functions in two variables, Barzilai and Borwein (1988) shows that the gradient method (1.10) with α_k given by (2.5) converges R-superlinearly and R-order is $\sqrt{2}$. To be more exact, we have the following result.

Theorem 2.1. (Barzilai-Borwein, 1988) *If $f(x)$ is a strictly quadratic convex function with 2 variables. The gradient method with BB step-size (2.5) almost always converges R-superlinearly in the sense that*

$$\|g_k\| \leq C\lambda^{-(\sqrt{2})^k} \tag{2.7}$$

holds asymptotically, where $\lambda = \sigma_1(H)/\sigma_2(H)$, C is a constant independent of k.

For the same objective function $f(x, y) = 100x^2 + y^2$ and the same starting point given in the previous section, the BB method produces the iterates as in the following picture (the first 9 iterations are given):



We can easily see that the BB method finds a very accurate solution after 9 iterations.

It is proved by Raydan (1993) that the BB method is global convergent for any n if the objective function is a convex quadratic. However, for $n > 2$, no superlinear convergence results have been established for the BB method, though numerical results indicates quite offen the BB method converges superlinearly.

As proved by Akaike (1959), the steepest descent method with exact line search will eventually reduce to a two-dimensional subspace spanned by two eigenvectors. This property, which was discovered by Dai and Fletcher (2003), is not possessed by the BB method. Actually, asymptotically the directions generated by the BB method spanned the whole space, namely the search directions of the BB method will not asymptotically converges to any lower dimensional subspace.

Due to the unexpected theoretical properties and the striking numerical performances of the BB method, it inspired lots of researches on the gradient methods, as such methods are widely used. In the next section, we will review the important advances on the different choices of the step-sizes α_k in the gradient method (1.10).

3 Some step-sizes for the gradient method

For simplicity, we assume that the objective function is the convex quadratic (1.5). It is easy to see that the Cauchy step-size is

$$\alpha_k^* = \frac{\|g_k\|_2^2}{g_k^T H g_k}. \tag{3.1}$$

It is interesting to notice that the BB step-size (2.6) is the Cauchy step-size in the previous iteration:

$$\alpha_k^{BB} = \frac{s_{k-1}^T s_{k-1}}{s_{k-1}^T y_{k-1}} = \alpha_{k-1}^*. \tag{3.2}$$

Because $\phi(\alpha) = f(x_k - \alpha g_k)$ is a convex quadratic function attaining its minimum at α_k^*, it is obvious that

$$\phi(\alpha) < \phi(0) \tag{3.3}$$

if and only if

$$\alpha \in (0, \quad 2\alpha_k^*). \tag{3.4}$$

The above condition is a necessary and sufficient condition for the step-size ensuring a reduction in the objective function, namely $f(x_{k+1}) < f(x_k)$.

For any step-size α_k, the gradient method (1.10) gives that

$$g_{k+1} = (I - \alpha_k H) g_k \tag{3.5}$$

which implies that

$$\|g_{k+1}\|_2 \leq \|I - \alpha_k H\|_2 \|g_k\|_2. \tag{3.6}$$

Minimizing the right hand side of the above inequality, Elman and Golub (1994) suggested that

$$\alpha_k^{OPT1} = \frac{2}{\lambda_1(H) + \lambda_n(H)}. \tag{3.7}$$

The step-size α_k^{OPT1} satisfies (3.4) because $\alpha_k^* \geq 1/\lambda_1(H)$. This step-size requires the estimates of the largest and smallest eigenvalues of H, which are generally not easy to obtain.

An approximate to the Cauchy step (3.1) is the following

$$\alpha_k^{OPT2} = \frac{\|g_k\|_2}{\|H g_k\|_2}, \tag{3.8}$$

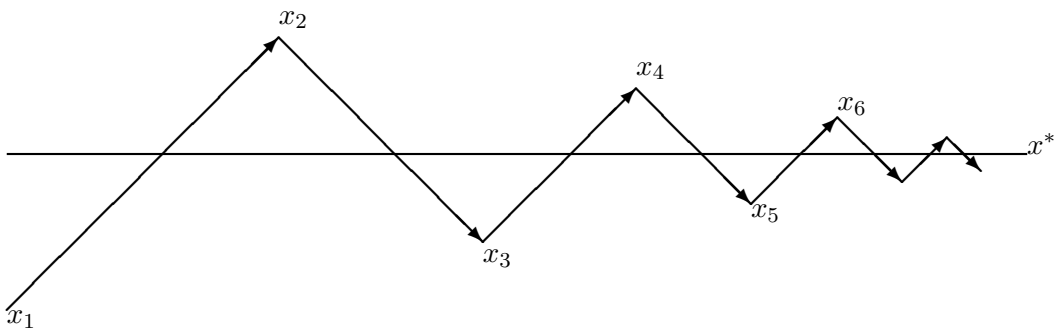
which was proposed by Dai and Yang (2001). They used the superscript "OPT2" as it turned out that this step-size will asymptotically converges to α_k^{OPT1}. Numerical results (Dai and Yang, 2001) indicate that both (3.7) and (3.8) behave very similar to the Cauchy step, and the search directions also converge to a two-dimensional subspace asymptotically.

One possible reason for the inefficiency of the Cauchy step (3.1) is that a constant criterion has been used in choosing the step-size, which leads a stable dynamic system giving slow convergence. To find a potential way to overcome this, Dai and Yuan (2003) proposed an alternate minimization (AM) gradient method, which minimizes the objective function and the norm of the gradient vector alternately. That is, the step-size is defined by

$$\alpha_k^{AM} = \begin{cases} \frac{g_k^T g_k}{g_k^T H g_k}, & \text{if } k \text{ is odd;} \\ \frac{g_k^T H g_k}{g_k^T H^2 g_k}, & \text{if } k \text{ is even.} \end{cases} \tag{3.9}$$

The alternately choice of step-sizes ensures the gradient method converging Q-linearly for n-dimensional quadratic problems, and it gives Q-superlinearly convergent for two-dimensional quadratic problems. Numerical results in Dai and Yuan (2003) show that (3.9) is significantly better than (3.1).

Another observation on the Cauchy step is that it is always too long in the sense that there exists a smaller step-size producing a gradient vector heading to the solution in two-dimensional case, which is illustrated in the following picture.



Thus, it is natural for us to consider step-sizes that are smaller than the Cauchy step. Actually, the AM step is a smaller step at the even iterations. Even though a smaller step can not achieve function reduction as much as the Cauchy step in the current iteration, it is more likely to reduce the function much more than the Cauchy step in the next iteration, which was observed for the AM method. Following this idea, Dai and Yuan (2003) suggested two shortened step-size methods

$$\alpha_k^{SS1} = \gamma_1 \alpha_k^* \tag{3.10}$$

and

$$\alpha_k^{SS2} = \begin{cases} \gamma_2 \alpha_k^*, & \text{if } k \text{ is odd;} \\ \alpha_k^*, & \text{if } k \text{ is even,} \end{cases} \tag{3.11}$$

where γ_1 and γ_2 are some positive constants less than 1. For example, we can let $\gamma_1 = 0.8$ and $\gamma_2 = 0.75$. Though the modifications are simple, the SS1 and SS2 methods avoid the zigzagging phenomenon to much extent and is comparable to the AM method. For any step-size α_k that satisfies

$$\delta_1 \le \frac{\alpha_k}{\alpha_k^*} \le \delta_2 \tag{3.12}$$

where $\delta_1 < \delta_2$ are two constants in (0,2), the convergence of the gradient method (1.10) can be similarly proved as for the Cauchy step. We can easily see that both SS1 and SS2 satisfies (3.12), hence it follows that both methods converge.

Raydan and Svaiter (2002) investigated the random choice of step-size α_k in the interval (3.4). Namely,

$$\alpha_k^{RAND} = \theta_k \alpha_k^*, \tag{3.13}$$

where θ_k is randomly chosen with a uniform distribution on $[0, 2]$. This random step-size gradient method also much outperforms the steepest descent method. Thus, it is surprising

to find that the "best" step-size ($\alpha_k = \alpha_k^*$, in the sense obtaining the least function value) is not as good as a random chosen step-size.

Hybrid methods can be constructed by using more than one of the step-sizes presented above. For example, Dai (2001) combined the Cauchy step and the BB step by suggesting the following alternate step method:

$$\alpha_k^{AS} = \begin{cases} \alpha_k^*, & \text{if } k \text{ is odd;} \\ \alpha_k^{BB}, & \text{if } k \text{ is even.} \end{cases} \tag{3.14}$$

In fact, the alternate step method is a method using the Cauchy step-size for every two consecutive iterations.

The above formulae for the step-sizes provide numerical improvements over the Cauchy step. However, it is undesirable that these improvements are not as great as what the BB method achieves, particularly when the condition number of the Hessian is very large. One consolation is that these step-sizes have the monotone property which is not inherited by the BB method. Thus, when we extend the BB method from convex quadratic minimization to general nonlinear unconstrained optimization, certain non-monotone techniques have to be applied, otherwise the fast convergence will be destroyed. Therefore, it is very desirable to find step-size formula which enables fast convergence and possesses the monotone property.

4 A new step-size

From the results of Akaike (1959), we see that the inefficiency of the Cauchy step is due to fact that the iterations will repeat(or cycle) the two iteration pattern. Therefore it is reasonable to believe that avoiding the iterates falling into a two dimensional subspace can escape from the inefficiency of the Cauchy step. Based on this belief, Yuan (2004) tried to find a step-size that would ensure finite termination for two dimensional quadratic problems. This goal can be achieved by defining

$$\alpha_k^Y = \frac{2}{\sqrt{(1/\alpha_{k-1}^* - 1/\alpha_k^*)^2 + 4\|g_k\|_2^2/\|\|s_{k-1}\|_2^2} + 1/\alpha_{k-1}^* + 1/\alpha_k^*}. \tag{4.1}$$

This step-size is defined in such a way that, for 2-dimensional convex quadratics problems, if

$$\alpha_1 = \alpha_1^*, \tag{4.2}$$
$$\alpha_2 = \alpha_2^Y, \tag{4.3}$$
$$\alpha_3 = \alpha_3^*, \tag{4.4}$$

then x_4 is the minimizer of the objective function in exact arithmetic. Based on this step-size, Yuan (2004) proposed two gradient algorithms, corresponding to the following

$$\alpha_k^{YA} = \begin{cases} \alpha_k^*, & \text{if } k \text{ is odd;} \\ \alpha_k^Y, & \text{if } k \text{ is even} \end{cases} \tag{4.5}$$

and

$$\alpha_k^{YB} = \begin{cases} \alpha_k^*, & \text{if } mod(k,3) \neq 0; \\ \alpha_k^Y, & \text{if } mod(k,3) = 0 \end{cases} \tag{4.6}$$

Both algorithms are monotone since it is easy to see from (4.1) that $\alpha_k^Y < 2\alpha_k^*$. The numerical experiments in Yuan (2004) show that (4.6) is comparable to the BB method for large scale problems and better for small scale problems. However, it is also found that (4.5) is far more worse than (4.6). The surprising numerical performance of (4.6) can be explained by the fact that (4.6) is nothing but repeated restarts of algorithm (4.2)-(4.4), which is proved to have the two-dimensional finite termination property.

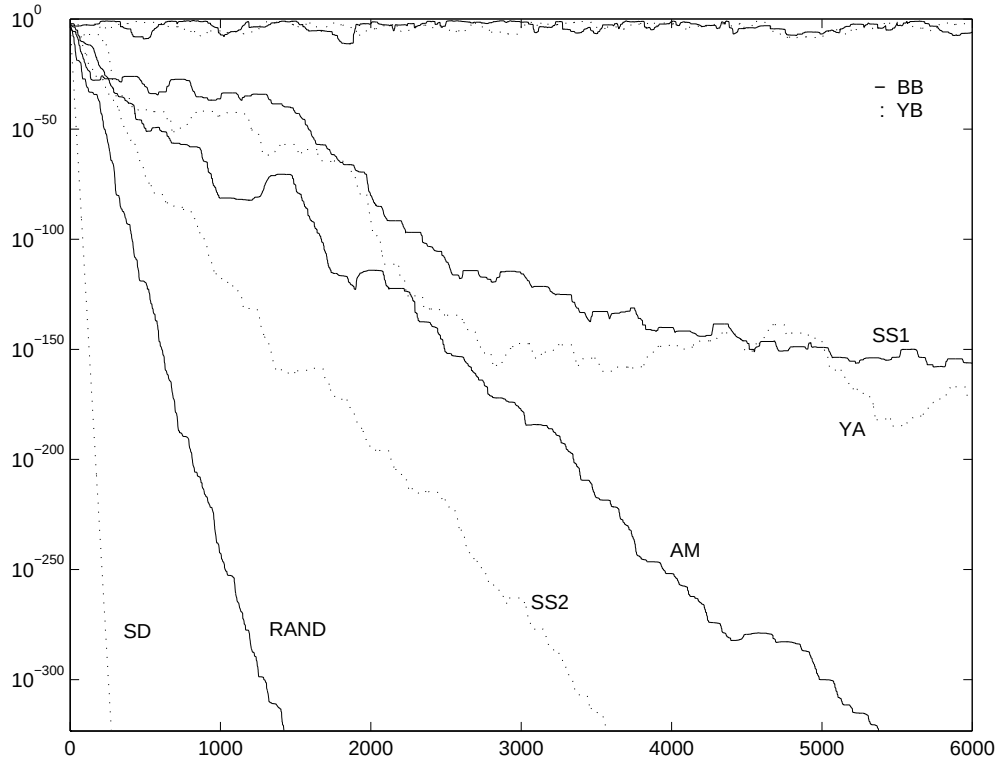
A distinguish feature of the BB method is that it does not sink into any lower subspace spanned by eigenvectors. To be more exact, the BB method reduces the gradient components (if they are expressed by the eigenvectors of H) more or less at the same asymptotic rate. This *decreasing together* property, as called by Dai and Yuan (2005), is also possessed by the method (4.6). Similar to Fletcher (2001), Dai and Yuan (2005) considers a special test example of 10 variables where H is a diagonal matrix with

$$h_{ii} = 11i - 10, \quad g_1^{(i)} = \sqrt{1+i}, \qquad \text{for } i = 1, \ldots, 10. \tag{4.7}$$

The quantity

$$\xi_k = \frac{\min\{\sum_{j=0}^{L-1} |g_{k+j}^{(i)}|;\ i = 1, \ldots, n\}}{\max\{\sum_{j=0}^{L-1} |g_{k+j}^{(i)}|;\ i = 1, \ldots, n\}} \tag{4.8}$$

is used to observe whether the gradient components decrease in a balanceable way. The introducing of a positive integer $L \geq 1$ is to smooth the curve of ξ_k with $L = 1$. It is set to 100 in our tests. The sequence $\{\xi_k\}$ generated by different gradient methods are given in the following picture.



The sequence $\{\xi_k\}$ generated by different gradient methods

From the picture, we can see that the method (4.6) shares with the BB method a property of decreasing together, which is out of the reach of the other gradient methods. In fact, it is observed that, except the BB method and the method (4.6), the other gradient methods converge asymptotically to lower dimensional subspaces. In fact, if we define the set

$$\mathcal{B} = \{i : 1 \leq i \leq n, \liminf_{k\to\infty} \frac{|g_k^{(i)}|}{\|g_k\|_2} \neq 0\} \tag{4.9}$$

and let n_b be the size of $\mathcal{B}$, namely, $n_b = |\mathcal{B}|$. Then for the previous example (4.7), Dai and Yuan (2005) obtained the following table.

Method	SD	RAND	SS1	SS2	AM	YA	YB	BB
n_b	2	2	3	2	4	7	10	10

The value n_b for different gradient methods in the example

The success of the step-size (4.1) leads to further investigation on similar step-sizes. Dai and Yuan (2005) considered the following variant step-size:

$$\alpha_k^{YV} = \frac{2}{\sqrt{(1/\alpha_{k-1}^* - 1/\alpha_k^*)^2 + 4\|g_k\|_2^2/(\alpha_{k-1}^*\|g_{k-1}\|_2)^2} + 1/\alpha_{k-1}^* + 1/\alpha_k^*}, \tag{4.10}$$

which is the same as (4.1) if $s_{k-1} = -\alpha_{k-1}^* g_{k-1}$. An extension to (4.1) given by Dai and Yuan(2005) is

$$\alpha_k^{Y2} = \frac{2}{\sqrt{(1/\alpha_{k-1}^{**} - 1/\alpha_k^{**})^2 + 4g_k^T H g_k/[(\alpha_{k-1}^{**})^2 g_{k-1}^T H g_k]} + 1/\alpha_{k-1}^{**} + 1/\alpha_k^{**}}, \tag{4.11}$$

where

$$\alpha_k^{**} = \frac{g_k^T H g_k}{\|H g_k\|_2^2}. \tag{4.12}$$

It is easy to show that the gradient method will find the minimum of a two dimensional quadratic function after three iterations if we let

$$\alpha_1 = \alpha_1^{**}, \tag{4.13}$$
$$\alpha_2 = \alpha_2^{Y2}, \tag{4.14}$$
$$\alpha_3 = \alpha_3^{**}. \tag{4.15}$$

Different combinations of either of (4.1), (4.10) and (4.11) with Cauchy steps are studied by Dai and Yuan (2005), and it was found that the following choice

$$\alpha_k^{DY} = \begin{cases} \alpha_k^*, & \text{if } mod(k,4) = 1 \text{ or } 2, \\ \alpha_k^{YV}, & \text{otherwise,} \end{cases} \tag{4.16}$$

produced better numerical results than the BB method.

5 Discussion

In this paper we have presented some new step-sizes for the gradient method and reviewed some of the recent advances, following the remarkable result by Barzilai and Borwein (1988). The step-sizes discussed in the paper show that there is still room for us to improve the classic steepest descent method. The new method (4.6) is interesting as it has the monotone property and converges as fast as the BB method. Numerical results indicate (4.6) converges Q-superlinearly for three-dimensional convex quadratic problems, but by now we have not yet managed to find a proof.

It is surprising to find that there is still much to be understood on one of the most simple unconstrained optimization methods, the steepest descent method. We believe that a good gradient method would use at least one exact line search (the Cauchy step) in every few iterations, as we do not want to miss the opportunity of finding the exact solution when the gradient at the current iterate point is an eigenvector of the Hessian. Another reason for supporting some Cauchy steps is that any superlinearly convergent step at an iteration in the gradient method would require that the step-size is close to the Cauchy step, asymptotically. Thus, it is natural to believe that a successful scheme would combine one or two Cauchy steps with one or two step-sizes defined by certain formulae. The formulae to be searched for would better have the following properties. First, it should be a monotone step so that it can be easily extended to general nonlinear optimization. Secondly, it should reduce the gradient components (expressed by the eigenvectors of the Hessian) more or less in the same rate. Finally, it should be easy to compute. For example, it should depend only on the information in the current iteration and the step-sizes used in the past few iterations.

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) 1-16.

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

[3] A. Cauchy, *Méthode générale pour la résolution des systèms d'equations simultanées*, Comp. Rend. Sci. Paris, 25 (1847), pp. 46-89.

[4] H.B. Curry, *The method of steepest descent for nonlinear minimization problems*, Quart. Appl. Math., 2 (1944) 258-261.

[5] Y.H. Dai, *Alternate step gradient method*, Report AMSS-2001-041, Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences, 2001.

[6] Y.H. Dai and R. Fletcher, *On the asymptotic behaviour of some new gradient methods*, Numerical Analysis Report, NA/212, Dept. of Math. University of Dundee, Scotland, UK (2003).

[7] Y. H. Dai and X. Q. Yang, *A New Gradient Method with an Optimal Stepsize Property*, Research report, Institute of Computational Mathematics and Scientific/Engineering Computing, Chinese Academy of Sciences, 2001.

[8] Y.H. Dai, J.Y. Yuan, and Y. Yuan, *Modified two-point step-size gradient methods for unconstrained optimization*, Computational Optimization and Applications, 22 (2002), 103-109.

[9] Y.H. Dai and Y. Yuan, *Alternate minimization gradient method*, IMA Journal of Numerical Analysis, 23 (2003) 377-393.

[10] Y.H. Dai and Y. Yuan, *Analysis of monotone gradient methods*, J. Industrial and Management Optimization, 1 (2005) 181-192.

[11] Y. H. Dai and H. Zhang, *An Adaptive Two-Point Step-size Gradient Method*, Research report, Institute of Computational Mathematics and Scientific/Engineering Computing, Chinese Academy of Sciences, 2001.

[12] H. C. Elman and G. H. Golub, *Inexact and preconditioned Uzawa algorithms for saddle point problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1645-1661.

[13] R. Fletcher, *Practical Methods of Optimization*(second Edition), John Wiley and Sons, Chichester, 1987.

[14] R. Fletcher, *On the Barzilar-Borwein method*, Research Report, University of Dundee, UK, 2001.

[15] G. E. Forsythe, *On the asymptotic directions of the s-dimensional optimum gradient method*, Numerische Mathematik, 11 (1968), pp. 57-76.

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

[17] J. Greenstadt, *On the relative efficiencies of gradient methods*, Math. Comp. 21 (1967) 360-367.

[18] J. Nocedal, A. Sartenaer and C. Zhu, *On the behavior of the gradient norm in the steepest descent method*, Research report, Northwestern University, USA, 2000

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

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

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

[22] E. Stiefel, *Über einige Methoden der Relaxationsrechnung*, Z. Angew. Math. Physik, 3 (1952) 1-33.

[23] M.N. Vrahatis, G.S. Androulakis, J.N. Lambrinos and G.D. Magoulas, *A class of gradient unconstrained minimization algorithms with adaptive step-size*, J. Comp. and Appl. Math. 114 (2000) 367-386.

[24] Y. Yuan, *A new stepsize for the steepest descent method*, Research report, Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences, 2004.