Relaxed Steepest Descent and Cauchy-Barzilai-Borwein Method

Marcos Raydan * Benar F. Svaiter †

January 3, 2001

Abstract

The negative gradient direction to find local minimizers has been associated with the classical steepest descent method which behaves poorly except for very well conditioned problems. We stress out that the poor behavior of the steepest descent methods is due to the optimal Cauchy choice of steplength and not to the choice of the search direction. We discuss over and under relaxation of the optimal steplength. In fact, we study and extend recent nonmonotone choices of steplength that significantly enhance the behavior of the method. For a new particular case (Cauchy-Barzilai-Borwein method), we present a convergence analysis and encouraging numerical results to illustrate the advantages of using nonmonotone overrelaxations of the gradient method.

Key Words: Steepest descent, gradient method with retards, Rayleigh quotient, Barzilai-Borwein method.

*Dpto. de Computación, Facultad de Ciencias, Universidad Central de Venezuela, Ap. 47002, Caracas 1041-A, Venezuela (mraydan@reacciun.ve). This author was partially supported by the Scientific Computing Center at UCV.

†Instituto de Matemática Pura e Aplicada, Estrada dona Castorina 110, Rio de Janeiro, RJ CEP 22460-320, Brazil (benar@impa.br). This work was partially supported by CNPq grant n. 301200/93-9(RN).

1 Introduction

The gradient direction has played an important role in the development of optimization techniques. Unfortunately, for the unconstrained minimization problem, it has also been associated with the classical and well-known steepest descent method [4]. This method has been frequently called the Cauchy method or simply the gradient method, and has been widely accepted that it converges rather slowly in most cases. The main goal of this work is to establish that the poor behavior of the Cauchy method is due to the optimal choice of steplength and not to the choice of the gradient direction.

In 1988, Barzilai and Borwein [2] presented a nonmonotone steplength associated with the gradient method that avoids the drawbacks of the Cauchy method. Later, Raydan [9] established global convergence in the convex quadratic case, and Dai and Liao [5] proved R-linear rate of convergence. Recently, Friedlander et al. [6] extended these results and presented a new family of nonmonotone gradient methods with retards. They establish convergence and illustrate with different examples the good behavior of these new methods.

We extend this line of research by studying the positive effects of using (over and under) relaxed steplengths of the gradient method for quadratics. In particular, we present an interesting member of the gradient method with retards family for which Q-linear rate of convergence can be established in a suitable norm. Each iteration of this version can be viewed as two consecutive steepest descent iterations in which the steplenght is computed once and used twice. As a consequence, the computational cost is similar to the one associated with the Cauchy method.

Our numerical experiments suggest that for the quadratic minimization problem, these new options clearly outperforms the classical Cauchy method and the Barzilai-Borwein method.

The rest of the paper is organized as follows. In section 2 we introduce and study the convergence of the relaxed Cauchy method. In section 3 we present numerical experiments to illustrate the behavior of randomly relaxed Cauchy methods when compared with the classical Cauchy method. In Section 4 the new Cauchy-Barzilai-Borwein method is presented, and its convergence as well as its computational cost are discussed. Finally, in Section 5 we present additional numerical results.

2 Relaxed Cauchy method

Our quadratic model problem is

$$\min f(x) = \frac{1}{2}x^t Q x - b^t x, \qquad x \in R^n \tag{1}$$

where $Q \in R^{n\times n}$ is symmetric positive definite (SPD) and $b \in R^n$. This problem is equivalent to solving the linear system:

$$Qx = b.$$

Since we are supposing Q to be positive definite, problem (1) has a unique solution given by $x^* = Q^{-1}b$.

The classical Cauchy method applied to problem (1) can be written as

$$x_{k+1} = x_k - \lambda_k g_k,$$

where $g_k = \nabla f(x_k) = Qx_k - b$ and the optimal choice of steplength λ_k is given by

$$\lambda_k = \frac{g_k^t g_k}{g_k^t Q g_k}.$$

It is well known that for the optimal choice λ_k the method possesses the following q-linear rate of convergence

$$\|x_k - x_*\|_Q \leq \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \|x_{k-1} - x_*\|_Q,$$

where for any $z \in \Re^n$, $\|z\|_Q^2 = z^t Q z$, and $\lambda_{\min}$ and $\lambda_{\max}$ represent the smallest and the largest eigenvalues of Q respectively. . This convergence rate result is obtained as a worst-case analysis. Nevertheless, in pratice, the asimptotic convergence rate is almost as bad as the one predicted by this analysis, making steepest descent a very slow method even for mild-conditioned problems. This phenomenon was explained by Akaike, in [1]. He proved that, unfortunately, the sequence of normalized errors $(x_k - x_*)/\|x_k - x_*\|$ accumulates in the two dimensional subspace generated by the two eigenvectors associated with $\lambda_{\min}$ and $\lambda_{\max}$.

Let us now modify the Cauchy method introducing over and under relaxation. Taking θ_k as relaxation parameters between 0 and 2 we get

$$x_{k+1} = x_k - \theta_k \lambda_k g_k, \tag{2}$$

where again $g_k = \nabla f(x_k) = Qx_k - b$ and $\lambda_k = \frac{g_k^t g_k}{g_k^t Q g_k}$. Notice that for $\theta_k = 1$ we obtain the classical steepest descent method, and also that for $\theta_k = 2$, $f(x_{k+1}) = f(x_k)$.

For any integer k the functional $a_k(x) = x^t Q^k x$ will be frequently used throughout this work. In particular, we now present some useful results (see also Brezinski [3] for additional results concerning totally monotonic sequences and Stieltjes moments).

Lemma 2.1 *For any $x \in \Re^n$, and for any $k \in Z$, it holds*

(i) The sequence $\{a_k(x)/a_{k-1}(x)\}$ is monotonically increasing.

(ii) $a_{k-2}(x)\ a_k(x) - a_{k-1}^2(x) \geq 0$.

(iii) $1 \geq a_{k-1}^2(x)/(a_k(x)\ a_{k-2}(x)) \geq 4\lambda_{min}\lambda_{max}/(\lambda_{min} + \lambda_{max})^2$,

where λ_{min} and λ_{max} are the smallest and largest eigenvalues of Q, respectively.

Proof. Since Q is SPD and $k - 1 = \frac{k}{2} + \frac{k-2}{2}$, we obtain using the Cauchy-Schwarz inequality

$$a_{k-1}^2(x) = ((Q^{k/2}x)^t (Q^{(k-2)/2}x))^2 \leq \|Q^{k/2}x\|_2^2\ \|Q^{(k-2)/2}x\|_2^2 = a_k(x)\ a_{k-2}(x),$$

which implies

$$a_k(x)/a_{k-1}(x) \geq a_{k-1}(x)/a_{k-2}(x).$$

Inequalities (ii) and the first part of (iii) follow directly from (i). The second part of (iii) is obtained by applying Kantorovich inequality to the matrix Q with the vector $y = \sqrt{Q^{k-1}}x$ (see [8]). □

The next theorem establishes convergence of the relaxed Cauchy method under very mild assumptions on the relaxation parameters θ_k.

Theorem 2.1 *If the sequence θ_k has an accumulation point $\bar{\theta} \in (0, 2)$ then x^k generated by the relaxed Cauchy method converges to x^*.*

Proof. Observe that for any k,

$$\phi_k(\theta) = f(x^k - \theta\lambda_k g_k).$$

is a second degree convex polinomial wich attains global minimum on $\theta = 1$. Hence by symmetry $\phi_k(0) = \phi_k(2)$ and for any $\theta \in [0, 2]$ $\phi_k(\theta) \leq \phi(0)$. Therefore

$$f(x^{k+1}) \leq f(x^k)$$

for all k. Since f is bounded below, then

$$\lim f(x^k) - f(x^{k+1}) = 0. \tag{3}$$

There exist some $\beta \in (0, 1)$ such that

$$\beta < \bar{\theta} < 2 - \beta$$

Consequently, there exist a subsequence θ_{k_j} contained in $[\beta, 2 - \beta]$. Using again the properties of ϕ_k we get $f(x^{k_j+1}) = \phi_{k_j}(\theta_{k_j}) < \phi_{k_j}(\beta)$. Since ϕ_k is convex, it follows that $\phi_{k_j}(\beta) \leq \beta(\phi_{k_j}(1) - \phi_{k_j}(1))$. By simple manipulations we obtain $\phi_k(1) - \phi_k(0) = (1/2)(g_k' g_k / g_k' Q g_k)$, wich combined with the previous equations yields

$$\begin{aligned} f(x^{k_j}) - f(x^{k_j+1}) &\geq (\beta/2)\frac{(g_{k_j}' g_{k_j})^2}{g_{k_j}' Q g_{k_j}} \\ &\geq \frac{\beta}{2\lambda_{\max}} \|g_{k_j}\|_2^2 \end{aligned} \tag{4}$$

Combining (3) and (4) we conclude that g_{k_j} goes to zero, and therefore x^{k_j} converges to x^*. Since $f(x^k)$ is nonincreasing, the whole sequence (of the functional values) converges to $f(x^*)$, which in turn implies convergence of x^k to x^*. $\square$

3 Randomly Relaxed Cauchy method

Theorem 2.1 opens interesting questions, for instance: Is it worth using the Cauchy method with relaxation? if yes, What are the good choices for the relaxation parameters?

In some particular cases the Cauchy choice of steplenght is the best possible choice. For example, if the search direction is an eigenvector, then clearly the Cauchy choice yields the global minimizer in one iteration. In those cases, the introduction of relaxation will not help. However, in practice this optimal situations happen very seldom, and relaxation might be a suitable tool to accelerate the convergence of the Cauchy method.

To illustrate the behavior of the Cauchy method introducing relaxation, we now present a numerical experiment where θ_k is chosen at random during the process, with a uniform distribution on $[0, 2]$. We report in Figure 1 the behavior of the Cauchy method and the random Cauchy method given by (2) when $f(x) = (1/2)x^tQx$ and the eigenvalues of Q are the positive integers from 1 to 1000, i.e., when the Euclidean condition number $\kappa_2(Q) = 1000$. Since the exact solution is given by the zero vector, we stop the process when the 2-norm of the iterates (error) is less than 10^{-12}. We observe that the random Cauchy method clearly outperforms the classical Cauchy method. We also observe that as predicted by Theorem 2.1 both methods converge monotonically to the unique minimizer. We run the same experiment several times with different random number generators obtaining similar results. This numerical experiment reveals the unfortunate and serious drawback that represents the use of the optimal Cauchy choice of steplength when searching the gradient direction.

4 The Cauchy-Barzilai-Borwein method

Motivated by the numerical experiment presented in the previous section, we go beyond in this section and discuss some gradient methods that use steplengths that do not guarantee descent in the objective function. In particular, we will consider methods for which $\theta_k > 2$ at some iterations, and the steplength is chosen by a prescribed formula.

The Barzilai-Borwein method, introduced in [2] for the unconstrained minimization problem can be written as

$$x_{k+1} = x_k - \frac{s_{k-1}^t s_{k-1}}{s_{k-1}^t (g_k - g_{k-1})} g_k,$$

where $s_{k-1} = x_k - x_{k-1}$. Notice that it requires, as well as the Cauchy method, only $O(n)$ floating point operations and one gradient evaluation per

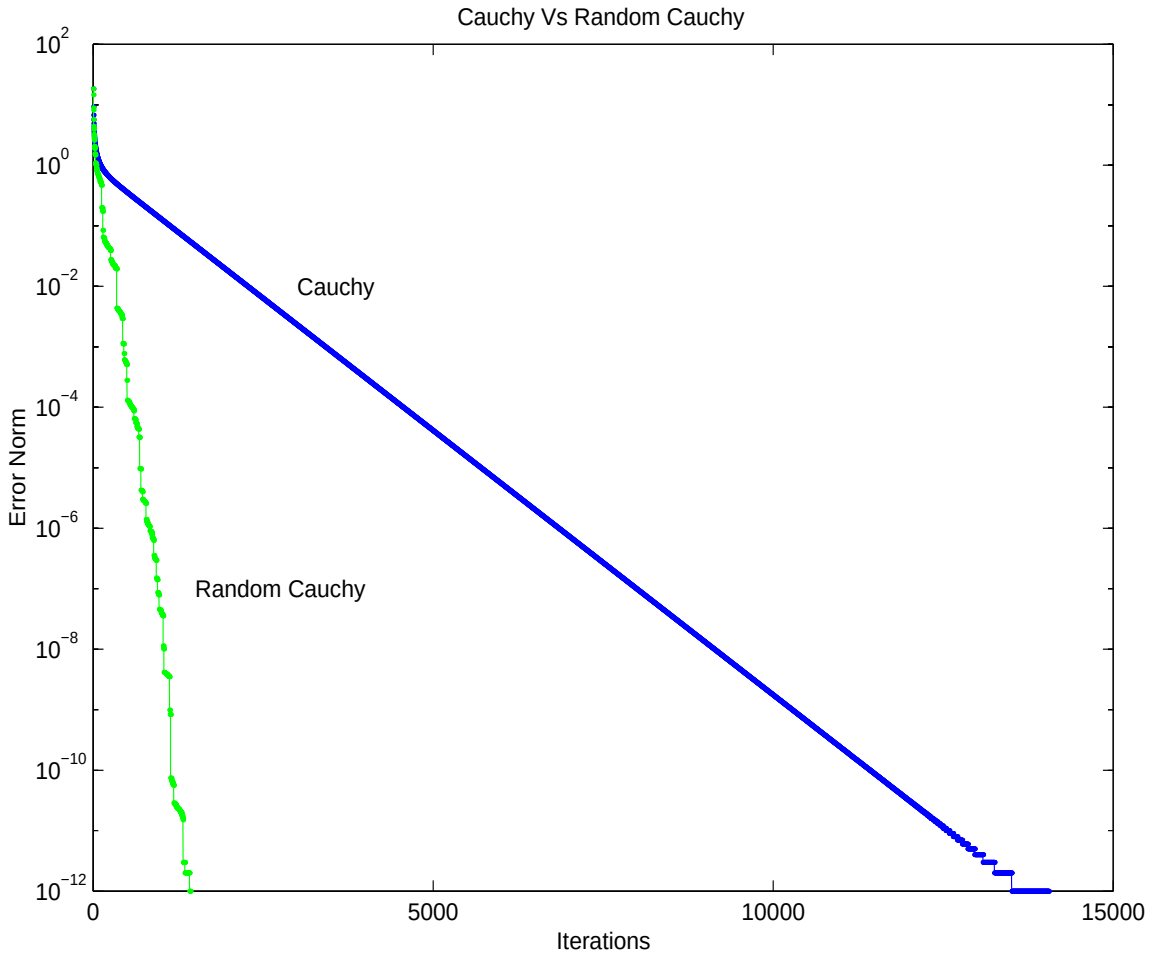

Figure 1: Cauchy Vs. random Cauchy when $\kappa_2(Q) = 1000$

iteration. The search direction is always the negative gradient direction but no line searches are required during the process. This inexpensive method greatly speeds up the convergence of gradient methods. See, for instance, the experiment reported in Figure 2 where we compare the behavior of the Barzilai-Borwein method and the random Cauchy method on the same experiment described in Section 3.

For quadratics, the Barzilai-Borwein method reduces to

$$x_{k+1} = x_k - \lambda_{k-1} g_k,$$

where λ_{k-1} is the optimal choice (Cauchy choice) at the previous iteration. Barzilai and Borwein [2] presented a convergence analysis for two dimensional convex quadratic problems. They also established, for that case, r-superlinear convergence. Later, Raydan [9] established global convergence for convex quadratic functions with any number of variables (see also [5]). Glunt, Hayden and Raydan [7] established a relationship with the shifted power method that adds understanding to the performance of this choice of steplength.

In this section, we propose a method which is in fact a modification of the Barzilai-Borwein method. At every other iterations, a Cauchy steplenght is evaluated once and used twice. So each pair of iterations can be done with almost the same computational cost of one Cauchy iteration. We will call this new nonmonotone method the Cauchy-Barzilai-Borwein (CBB) method, which is described by the following algorithm:

Take x_0 in R^n, and at every iteration k, do

$$\begin{aligned} g_k &= \nabla f(x_k) \\ &= Qx_k - b, \end{aligned}$$

$$h_k = Qg_k,$$

$$t_k = ({g_k}^t g_k)/({g_k}^t h_k),$$

$$y_k = x_k - t_k g_k,$$

$$\begin{aligned} x_{k+1} &= y_k - t_k \nabla f(y_k) \\ &= y_k - t_k(Qy_k - b). \end{aligned}$$

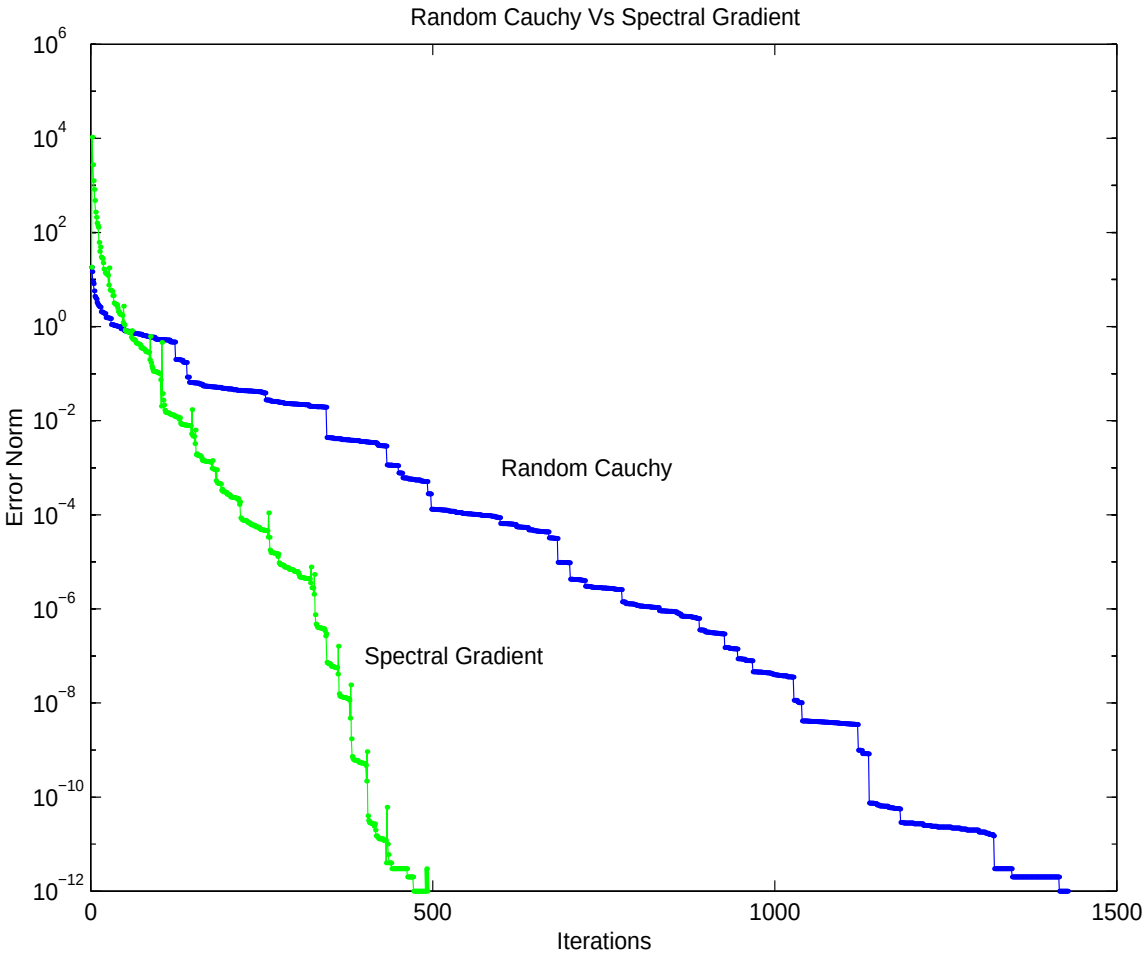

Figure 2: Barzilai-Borwein Vs. random Cauchy when $\kappa_2(Q) = 1000$

Note that

$$\begin{aligned} Qy_k - b &= Q(x_k - t_k g_k) - b \\ &= Qx_k - b - t_k Q g_k \\ &= g_k - t_k h_k. \end{aligned}$$

Therefore

$$\begin{aligned} x_{k+1} &= y_k - t_k(g_k - t_k h_k) \\ &= x_k - 2t_k g_k + {t_k}^2 h_k. \end{aligned} \tag{5}$$

The computationally efficient version of the iterative step is given by:

$$g_k = Qx_k - b, \quad h_k = Qg_k,$$

$$t_k = {g_k}^t g_k / ({g_k}^t h_k),$$

$$x_{k+1} = x_k - 2t_k g_k + {t_k}^2 h_k.$$

Convergence of the sequence $\{x_k\}$ to x^* can be proved using the results of Friedlander et al. [6], developed in a more general framework. One can easly observe that the sequences $\{||x_k - x_*||\}$ and $\{f(x_k) - f(x^*)\}$, although converging to zero, are non-monotonically decreasing. Nevertheless, extensive numerical experiments show that algorithm CBB is much more efficient than the classical Cauchy method.

Some observations are in order: The number of matrix vector multiplications per iteration in algorithm CBB is the same as in the Cauchy method, i.e., 2. (one to evaluate g and one to evaluate h) Algorithm CBB performs one more vector sum and one more inner product than the Cauchy method.

The comparison of our algorithm with the Barzilai-Borwein (BB) method gives the folowing: If we count each iteration of algorithm CBB as two (evaluating y_k and then x_{k+1}) then the mean convergence rate of these methods are roughly the same, and the computational work of CBB is almost one half of the BB, because in each two cycles, g and h are evaluated only once.

If we count the iterations of CBB as in our definition, the mean convergence of the algorithm CBB is twice the mean convergence of BB and the computational cost per iteration is sligthly higher, (one more vector sum and one more scalar vector multiplication per iteration)

We will now prove that the sequence $\{x_k\}$ converges (Q–linearly) to x^* in the elliptic norm $||\cdot||_{Q^{-1}}$ defined by

$$||x||_{Q^{-1}} = \sqrt{x^t Q^{-1} x}$$

Although this norm is not suitable for practical purposes, it is suitable for theoretical reasons. This norm is also induced by the inner-produt $< \cdot, \cdot >_{Q^{-1}}$ defined by

$$< x, y >_{Q^{-1}} = x^t Q^{-1} y,$$

and it satisfyes the Cauchy-Schwarz inequality:

$$(||x||_{Q^{-1}})^2 (||y||_{Q^{-1}})^2 \geq < x, y >^2_{Q^{-1}} .$$

Theorem 4.1 *The sequence $\{x_k\}$ generated by the CBB method (5) converges Q-linearly in the norm $\|\cdot\|_{Q^{-1}}$ with convergence factor*

$$1 - \theta = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{max}}.$$

Proof. First observe that

$$x_{k+1} - x^* = (I - t_k Q)^2 (x_k - x^*).$$

Therefore,

$$(||x_{k+1} - x^*||_{Q^{-1}})^2 = (x_k - x^*)^t (I - t_k Q)^2 Q^{-1} (I - t_k Q)^2 (x_k - x^*).$$

To simplify the calculations of the proof, define for $x \in R^n$, $p \in Z$:

$$b_p(x) = (x - x^*)^t Q^p (x - x^*).$$

Note that for $p \geq 2$, $b_p(x)$ can be evaluated without knowing x^* or solving linear systems with Q, by means of $\nabla f(x)$:

$$b_p(x) = \nabla f(x)^t Q^{p-2} \nabla f(x).$$

(We shall use b_p insted of $b_p(x)$ when there is no doubt about the x.)

Observe that

$$||x - x^*||^2_{Q^{-1}} = b_{-1}(x),$$

and so

$$(||x_k - x^*||_{Q^{-1}})^2 = b_{-1}(x_k).$$

Combining with (5) yields

$$b_{-1}(x_{k+1}) = b_{-1}(x_k) - 4t_k b_0(x_k) + 6{t_k}^2 b_1(x_k) - 4{t_k}^3 b_2(x_k) + {t_k}^4 b_3(x_k).$$

To simplify the notation, let us set

$$b_p = b_p(x_k).$$

Then, we obtain

$$\begin{aligned} b_{-1}(x_{k+1}) &= b_{-1} - 4t_k b_0 + 6{t_k}^2 b_1 - 4{t_k}^3 b_2(x_k) + {t_k}^4 b_3(x_k) \\ &= b_{-1} - t_k(4b_0 - 6t_k b_1 + 4{t_k}^2 b_2 - {t_k}^3 b_3) \\ &= b_{-1} - t_k(4b_0 - 6t_k b_1 + 3{t_k}^2 b_2 + {t_k}^2(b_2 - t_k b_3)). \end{aligned}$$

Note that

$$b_2 = b_2(x_k) = {g_k}^t g_k,$$

and

$$b_3 = b_3(x_k) = {g_k}^t h_k.$$

Therefore

$$t_k = b_2/b_3,$$

Using Lemma 2.1, we obtain that the term $b_2 - t_k b_3$ vanishes and we get:

$$\begin{aligned} b_{-1}(x_{k+1}) &= b_{-1} - t_k(4b_0 - 6t_k b_1 + 3{t_k}^2 b_2) \\ &= b_{-1} - t_k(4b_0 - 6t_k b_1(x_k) + 3{t_k}^2 b_2) \\ &= b_{-1} - t_k(b_0 + 3b_0 - 6t_k b_1 + 3{t_k}^2 b_2) \\ &= b_{-1} - t_k(b_0 + 3(b_0 - 2t_k b_1 + {t_k}^2 b_2)) \end{aligned}$$

The second-degree polinomial $b_0 - 2tb_1 + t^2 b_2$ attains its minimum at $\hat{t} = b_1/b_2$. At this point, its value is

$$b_0 - b_1^2/b_2 = (b_0 b_2 - b_1^2)/b_2.$$

Direct calculations give:

$$b_0 b_2 - b_1^2 = ||x_k - x^*||^2 ||Q(x_k - x^*)||^2 - (x_k - x^*)' Q(x_k - x^*)$$

Therefore, directly from the Cauchy-Schwarz inequality we conclude that

$$b_0 - b_1^2/b_2 = (b_0 b_2 - b_1^2)/b_2 \geq 0.$$

Expanding the above polinomial aroud $\hat{t}$ we get

$$b_0 - 2tb_1 + t^2 b_2 = (b_0 b_2 - b_1^2)/b_2 + b_2(t - \hat{t})^2.$$

Hence

$$\begin{aligned} b_{-1}(x_{k+1}) &= b_{-1} - t_k \left[b_0 + 3 \left(\frac{b_0 b_2 - b_1^2}{b_2} + b_2 (t_k - \hat{t})^2 \right) \right] \\ &= b_{-1} - t_k (b_0 + 3((b_0 b_2 - b_1^2)/b_2 + b_2(b_2/b_3 - b_1/b_2)^2), \end{aligned}$$

and we conclude that

$$b_{-1}(x_{k+1}) = b_{-1}(1 - \theta_k),$$

where

$$\begin{aligned} \theta_k &= b_0 b_2/(b_{-1} b_3) + \tfrac{3}{(b_{-1} b_3)}(b_0 b_2 - b_1^2) \\ &\quad + \tfrac{3 b_2^2}{(b_{-1} b_3)}(b_1/b_2 - b_2/b_3)^2. \end{aligned} \tag{6}$$

Finally, using (6), Lemma 2.1, and properties of the Rayleigh quotient, we obtain

$$\theta_k \geq b_0 b_2/(b_{-1} b_3) \geq \lambda_{min}/\lambda_{max} > 0,$$

and the result is established. □

5 Numerical Experiments

To illustrate the behavior of the CBB method, we compare in Figure 3 the CBB with all previous methods on the same experiment described in Section 3. We observe that the CBB method outperforms the Barzilai-Borwein and also the random Cauchy method. We run the same experiment several times with different random number generators obtaining similar results.

In the same experiment we also report with red circles the iterations in which the search direction g_k is *almost* an eigenvector of the matrix Q. To do so we check if g_k is *almost* parallel to Qg_k, i.e., if

$$cos\ (g_k, Qg_k) = \frac{g_k^t Q g_k}{||g_k||_2\ ||Qg_k||_2} > 1 - \epsilon, \tag{7}$$

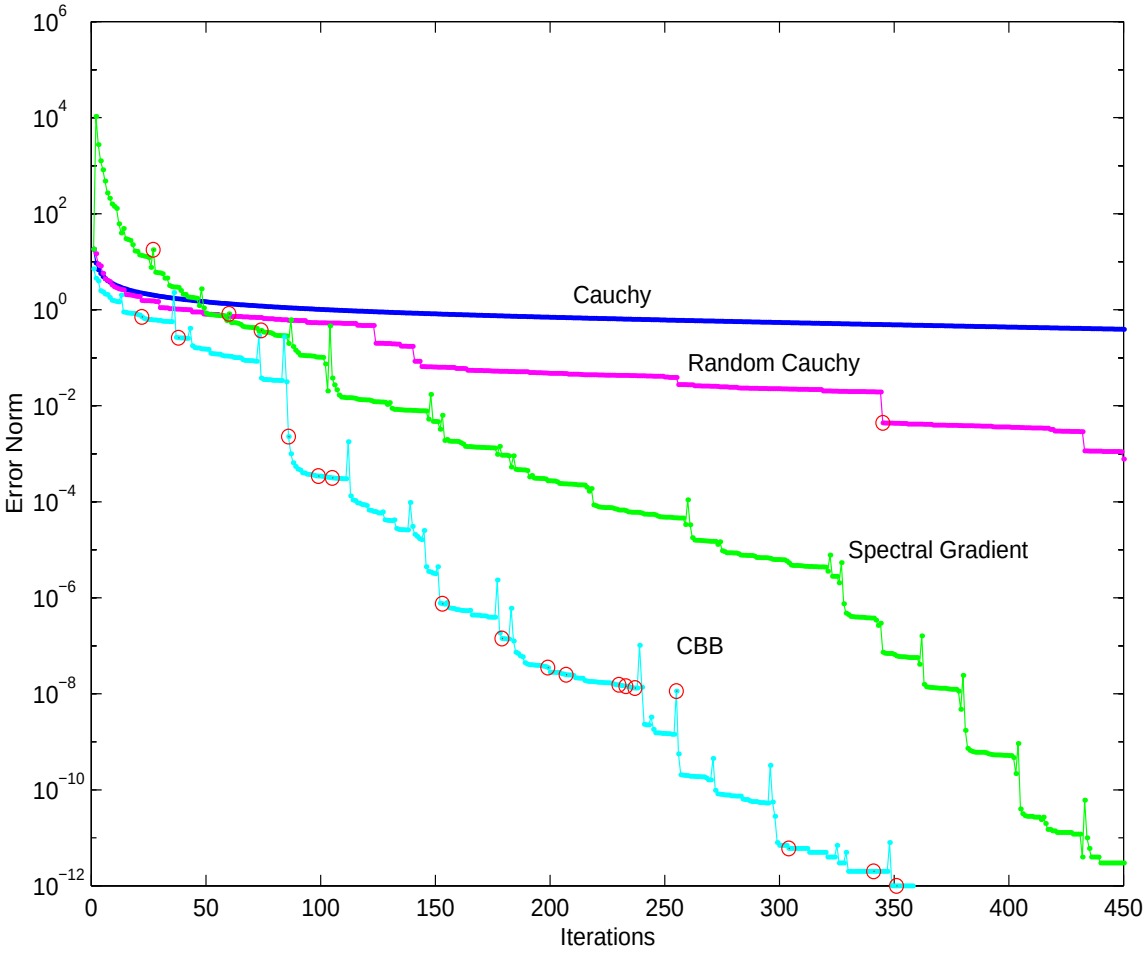

Figure 3: Number of iterations of all methods when $\kappa_2(Q) = 1000$

n	Cauchy	random Cauchy	Barzilai-Borwein	CBB
50	813	315	108	79
500	8003	916	402	230
1000	17053	2003	517	392

Table 1: Average number of iterations on 5 random experiments

n	Cauchy	random Cauchy	Barzilai-Borwein	CBB
50	0	2	3	5
500	0	5	6	14
1000	0	6	8	18

Table 2: Average number of iterations at which g_k is *almost* an eigenvector on 5 random experiments

where $\epsilon > 0$ is small. For our experiments we choose $\epsilon = 0.0005$. This test is checked until convergence.

We also observe from Figure 3 that the choice of steplength of the CBB method tends to force gradient directions that approximates eigenvectors of the Hessian matrix Q. The same tendency is observed, but not as frequently, in the Barzilai-Borwein method. The approximation of eigenvectors during the process is a nice feature that explains the good behavior and the acceleration observed for these nonmonotone gradient methods (see [6] and [7] for a relationship between the gradient methods with retards and the shifted power method to approximate eigenvectors).

Finally, we report on tables 1 and 2 the average number of iterations for convergence, and the average number of iterations at which g_k is *almost* an eigenvector (satisfies (7)) on 5 random experiments for which $\kappa_2(Q) = n$, for different values of n. Once again we oberve the superiority of CBB over all other methods, and we also observe the previously described tendency of the nonmonotone methods to approximate eigenvectors during the process.

Aknowledgements. We would like to thank Alejandra Alonzo and Ana Luis from Universidad Central de Venezuela for programming assistance in obtaining our numerical results.

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*, 11:1–16, 1959.

[2] J. BARZILAI and J.M. BORWEIN. Two point step size gradient methods. *IMA J. Numer. Anal.*, 8:141–148, 1988.

[3] C. BREZINSKI. *Padé–Type Approximation and General Orthogonal Polynomials*. Birkhauser–Verlag, Basel, 1980.

[4] A. CAUCHY. Méthodes générales pour la résolution des systèmes d'équations simultanées. *C. R. Acad. Sci. Par.*, 25:536–538, 1847.

[5] Y. H. DAI and L. Z. LIAO. R-linear convergence of the Barzilai and Borwein gradient method. Technical Report AMSS 1999-081, Academy of Mathematics and Systems Sciences, Beijing, China, 1999.

[6] A. FRIEDLANDER, J.M. MARTINEZ, B. MOLINA, and M. RAYDAN. Gradient method with retards and generalizations. *SIAM J. Numer. Anal.*, 36:275–289, 1999.

[7] W. GLUNT, T.L. HAYDEN, and M. RAYDAN. Molecular conformations from distance matrices. *J. Comp. Chem.*, 14:114–120, 1993.

[8] D. LUENBERGER. *Linear and Nonlinear Programming*. Addison-Wesley, Menlo Park, CA, 1984.

[9] M. RAYDAN. On the Barzilai and Borwein choice of steplength for the gradient method. *IMA J. Numer. Anal.*, 13:321–326, 1993.