ON THE CONVERGENCE RATE OF THE CAUCHY ALGORITHM IN THE l_2 NORM

ELIZABETH W. KARAS [1]
ALESSANDRA M. DA MOTA[2] ADEMIR A. RIBEIRO[1]

May 9, 2005

[1]Department of Mathematics, Federal University of Paraná. Cx. Postal 19081, 81531-980 Curitiba, PR, Brazil; e-mail :karas@mat.ufpr.br, ademir@mat.ufpr.br.

[2]Msc in Numerical Methods-CESEC, Federal University of Paraná. Cx. Postal 19081, 81531-980 Curitiba, PR, Brazil; e-mail : ammota@uepg.br.

Abstract. This paper presents a convergence rate for the sequence generated by the Cauchy algorithm. The method is applied to a convex quadratic function with exact line search. Instead of using the norm induced by the hessian matrix, the q-linear convergence is shown for the l_2 (or Euclidean) norm.

Key words. Cauchy method, steepest descent, q-linear convergence.

1. Introduction

The Steepest Descent or Cauchy algorithm is one of the oldest and most widely known methods for minimizing a function of several variables.

When the Armijo search is performed it is well known that the algorithm is globally convergent, in the sense that all limit points of the generated sequence are critical. This result is proved in many nonlinear programming textbooks, see Ref. 2, 3, 14.

Under convexity, it has been shown Ref. 6, 11, 12 that the sequence produced by the algorithm with Armijo line searches is actually convergent whenever there are optimal solutions. This is not true if we use exact line search. In Ref. 11 we have an example of a convex continuously differentiable function for which the Cauchy algorithm generates four limit points.

Regarding speed of convergence, if exact line search is used, we have the classical result that ensures, under usual hypotheses, q-linear convergence for the sequence of objective values. The convergence rate is bounded above by $((\kappa-1)/(\kappa+1))^2$, where κ is the condition number of the hessian at the solution, Ref. 2, 3, 13. When the function is quadratic and convex, this means that the sequence (x^k) converges q-linearly, with the norm induced by the hessian, at rate $(\kappa-1)/(\kappa+1)$ (see Lemma 3.1 bellow). As we know, q-linear convergence is a norm-dependent property, and it does not follow from this that the sequence (x^k) converges q-linearly in l_2 norm. In Section 2. we present a sequence for which the q-linear convergence is lost when we change the norm.

We have also other choices for the stepsize. In the Barzilai-Borwein method Ref. 1, it is taken as the standard exact line search stepsize, but using the gradient vector of the preceding iteration, instead of the current one. This choice has an unexpected result for the steepest descent method, R-superlinear convergence for a quadratic and convex function in $\mathbb{R}^2$, for

almost all the starting points. We can find further discussions about this method in Ref. 4, 5, 7, 8, 9, 15.

In Ref. 3 we find a result that establishes the best convergence rate for a constant stepsize: for a convex quadratic function it holds that

$$\frac{\|x^{k+1} - x^*\|_2}{\|x^k - x^*\|_2} \leq \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}, \tag{1}$$

with the stepsize

$$\alpha = \frac{2}{\lambda_n + \lambda_1}. \tag{2}$$

Here λ_1 and λ_n are the smallest and largest eigenvalues of the Hessian, respectively.

Another interesting report on the Cauchy algorithm can be found in Ref. 10, where a new stepsize is proposed. This stepsize approaches the "optimal" stepsize (2) as $k \to \infty$, so that (1) is achieved asymptotically.

By the way, as far as we could verify, there are few references that consider the convergence rate of (x^k) in l_2 norm. This was one motivation for the study presented here. The other motivation came from the classical picture in Figure 1 that shows a decreasing sequence of Euclidean distances from the iterates to the minimizer.

The inequality in (1) suggests a natural question. Does this bound hold for the exact line search? The answer is no. We give in Section 4. an example for which this bound does not hold. Nevertheless, we have shown that the decrease of $\left(\|x^k - x^*\|_2\right)$ has a linear bound when the Cauchy method is applied to a convex quadratic function with exact line search. Furthermore the convergence rate is bounded above by

$$\sqrt{1 - \frac{\lambda_1}{\lambda_n}} = \sqrt{1 - \frac{1}{\kappa}},$$

where κ is the condition number of the hessian. We prove this in Section 3..

This paper is organized as follows. In Section 2. we present an example showing the norm dependence of q-linear convergence. In Section 3. we prove the q-linear convergence of the Cauchy algorithm in the l_2 norm. Section 4. presents some discussions about the bound given in this paper and the standard bound given in (1).

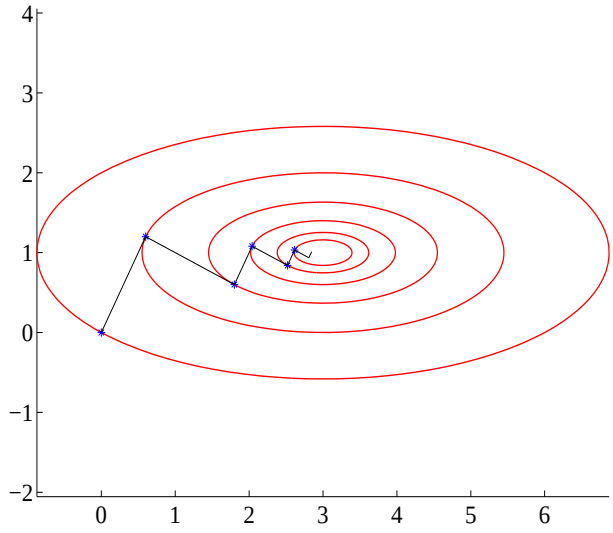

Figure 1: The path of Cauchy algorithm.

2. The norm dependence of q-linear convergence

We know that convergence of a sequence (x^k) in $\mathbb{R}^n$ does not depend on norm, because the norms in $\mathbb{R}^n$ are equivalent. However, there exist sequences that converge q-linearly in one norm but not in the other. Let us see such an example.

First we introduce, for a given positive definite matrix Q in $\mathbb{R}^{n\times n}$, the induced norm

$$\|x\|_Q = \sqrt{x^T Q x}. \tag{3}$$

Consider the sequence (x^k) in $\mathbb{R}^2$ given by

$$x^{2k} = \begin{pmatrix} \dfrac{1}{(2\sqrt{2})^k} & 0 \end{pmatrix}^T \qquad \text{and} \qquad x^{2k+1} = \frac{1}{(2\sqrt{2})^k} \begin{pmatrix} \dfrac{1}{\sqrt{2}} & \dfrac{1}{\sqrt{2}} \end{pmatrix}^T.$$

Define the ratio $r_k = \dfrac{\|x^{k+1}\|_Q^2}{\|x^k\|_Q^2}$, where $Q = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{4} \end{pmatrix}$. It can be easily seen that $r_{2k} = 5/8$ and $r_{2k+1} = 1/5$, for all $k \in \mathbb{N}$. Hence the sequence (x^k) converges q-linearly to zero, with the norm $\|\cdot\|_Q$.

Now consider the ratio $s_k = \|x^{k+1}\|_2^2/\|x^k\|_2^2$. In this case we have $s_{2k} = 1$ for all $k \in \mathbb{N}$. Therefore (x^k) does not converge q-linearly in l_2. Figure 2 shows the sequence and the contours of the norms considered here.

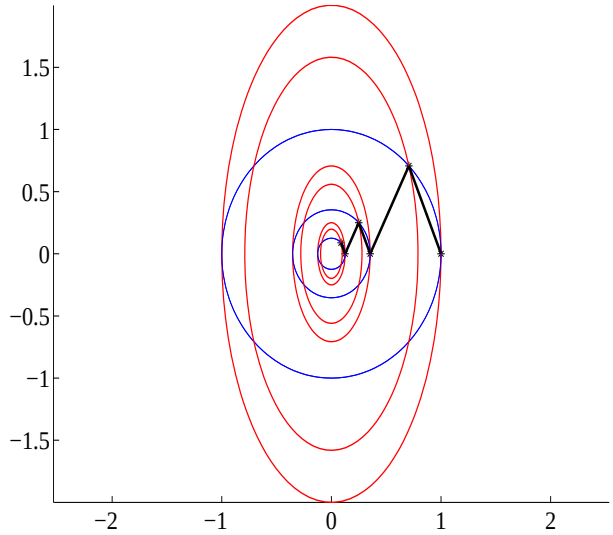

Figure 2: The norm dependence of q-linear convergence.

3. The q-linear convergence in the l_2 norm

Consider the problem of unconstrained minimization of the quadratic function

$$f(x) = \frac{1}{2}x^T Q x + b^T x + c, \tag{4}$$

where Q is a positive definite matrix in $\mathbb{R}^{n\times n}$. In this simple case, the unique minimizer of f can be found directly. It is the point x^* satisfying

$$Qx^* + b = 0. \tag{5}$$

We consider the Cauchy algorithm with exact line search to obtain this minimizer.

Algorithm 3.1. Cauchy

Data: $k = 0$, $x^0 \in \mathbb{R}^n$, $d^0 = -\nabla f(x^0)$,
WHILE $d^k \neq 0$
  Compute the stepsize α_k such that $f(x^k + \alpha_k d^k) = \min_{\alpha \geq 0} f(x^k + \alpha d^k)$
  $x^{k+1} = x^k + \alpha_k d^k$
  $k = k + 1$
  $d^k = -\nabla f(x^k)$
END.

In this particular case we can determine the stepsize α_k explicitly. By differentiating the function $\alpha \in \mathbb{R} \mapsto f(x^k + \alpha d^k)$ and setting its derivative to zero, we obtain

$$\alpha_k = \frac{(d^k)^T d^k}{(d^k)^T Q d^k}. \tag{6}$$

The next lemma shows that q-linear convergence for $(f(x^k))$ implies the q-linear convergence for (x^k) with the induced norm $\|\cdot\|_Q$ given by (3). (They are, in fact, equivalent).

Lemma 3.1. Let (x^k) be any sequence in $\mathbb{R}^n$. Suppose that the sequence $(f(x^k))$ satisfies $f(x^{k+1}) - f(x^*) \leq \beta(f(x^k) - f(x^*))$, where $0 \leq \beta < 1$ and x^* is given by (5). Then

$$\|x^{k+1} - x^*\|_Q \leq \sqrt{\beta}\, \|x^k - x^*\|_Q.$$

Proof. Using (5), for all $x \in \mathbb{R}^n$ we have

$$\begin{aligned} \tfrac{1}{2}\|x - x^*\|_Q^2 &= \tfrac{1}{2}(x - x^*)^T Q (x - x^*) \\ &= \tfrac{1}{2} x^T Q x - x^T Q x^* + \tfrac{1}{2} x^{*T} Q x^* \\ &= \tfrac{1}{2} x^T Q x + b^T x + \tfrac{1}{2} x^{*T} Q x^*. \end{aligned}$$

But, again using (5), $f(x^*) = -\frac{1}{2} x^{*T} Q x^* + c$. Thus

$$\frac{1}{2}\|x - x^*\|_Q^2 = f(x) - f(x^*).$$

Therefore

$$\begin{aligned} \|x^{k+1} - x^*\|_Q &= \sqrt{2(f(x^{k+1}) - f(x^*))} \\ &\leq \sqrt{2\beta(f(x^k) - f(x^*))} \\ &= \sqrt{\beta}\, \|x^k - x^*\|_Q, \end{aligned}$$

completing the proof. □

As we discuss in Section 2., despite the fact that two norms in $\mathbb{R}^n$ are equivalent, we cannot conclude from Lemma 3.1 that (x^k) has q-linear convergence with the Euclidean norm. However, this result is actually true and we give now a direct proof.

We may assume without loss of generality that $x^* = 0$ and $f(x^*) = 0$, that is

$$f(x) = \frac{1}{2} x^T Q x, \tag{7}$$

otherwise consider the function $g(x) = f(x + x^*) - f(x^*) = \frac{1}{2} x^T Q x$.

Let us denote the eigenvalues of the hessian Q by

$$0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n.$$

Lemma 3.2. Given $x \in \mathbb{R}^n$, $x \neq 0$, set $d = -Qx$ and $\alpha = (d^T d)/(d^T Q d)$. Then

$$\alpha \leq \frac{x^T Q x}{x^T Q^2 x}.$$

Proof. Since $d = -Qx$, we have $x^T Q x = d^T Q^{-1} d$ and $x^T Q^2 x = d^T d$. Hence

$$\alpha \frac{x^T Q^2 x}{x^T Q x} = \frac{(d^T d)^2}{(d^T Q d)(d^T Q^{-1} d)}. \tag{8}$$

Since Q is a positive definite matrix, we can write $Q = GG^T$, for some matrix $G \in \mathbb{R}^{n \times n}$. By making $u = G^T d$ and $v = G^{-1} d$, we have $u^T v = d^T d$, $u^T u = d^T Q d$ and $v^T v = d^T Q^{-1} d$. By the Cauchy-Schwarz inequality, we conclude from (8) that $\alpha(x^T Q^2 x) \leq (x^T Q x)$, completing the proof. □

Theorem 3.1. Consider the quadratic function given in (7) and the sequence (x^k) generated by the Cauchy algorithm 3.1. Then, for all $k \in \mathbb{N}$,

$$\|x^{k+1}\|_2 \leq \gamma \|x^k\|_2,$$

where $\gamma = \sqrt{1 - \lambda_1/\lambda_n}$.

Proof. We have $d^k = -\nabla f(x^k) = -Qx^k$, so

$$\begin{aligned} \|x^{k+1}\|_2^2 &= (x^{k+1})^T x^{k+1} \\ &= (x^k + \alpha_k d^k)^T (x^k + \alpha_k d^k) \\ &= (x^k)^T x^k + 2\alpha_k (x^k)^T d^k + \alpha_k^2 (d^k)^T d^k \\ &= \|x^k\|_2^2 - 2\alpha_k (x^k)^T Q x^k + \alpha_k^2 (x^k)^T Q^2 x^k. \end{aligned}$$

From Lemma 3.2,

$$\|x^{k+1}\|_2^2 \leq \|x^k\|_2^2 - 2\alpha_k (x^k)^T Q x^k + \alpha_k (x^k)^T Q x^k = \|x^k\|_2^2 - \alpha_k (x^k)^T Q x^k.$$

If $x^k = 0$ there is nothing to do. So, suppose that $x^k \neq 0$. Using (6), we obtain

$$\frac{\|x^{k+1}\|_2^2}{\|x^k\|_2^2} \leq 1 - \frac{(d^k)^T d^k}{(d^k)^T Q d^k} \frac{(x^k)^T Q x^k}{(x^k)^T x^k}. \tag{9}$$

From Ref. 3, Proposition A.18 we have

$$\frac{(d^k)^T d^k}{(d^k)^T Q d^k} \geq \frac{1}{\lambda_n} \qquad \text{and} \qquad \frac{(x^k)^T Q x^k}{(x^k)^T x^k} \geq \lambda_1.$$

By substituing these inequalities in (9), it follows that

$$\frac{\|x^{k+1}\|_2^2}{\|x^k\|_2^2} \leq 1 - \frac{\lambda_1}{\lambda_n},$$

completing the proof. □

This theorem has an interesting geometrical interpretation. The level curves of f are ellipsoids whose eccentricity depends on the distance between the smallest and largest eigenvalues of the hessian Q. If $\lambda_1 = \lambda_n$, then the level curves are spheres and convergence occurs in one step. However, if $\lambda_1 \ll \lambda_n$, then the ellipsoids become very eccentric and the convergence rate is slowed. Note that, even if all but one eigenvalues are equal and the other is far away, the effectiveness of convergence is destroyed.

4. Discussion

In this paper, we have proved that

$$\frac{\|x^{k+1} - x^*\|_2}{\|x^k - x^*\|_2} \leq \sqrt{1 - \frac{\lambda_1}{\lambda_n}}, \tag{10}$$

when the Cauchy method is applied to a convex quadratic function with exact line search. As we comment in Section 1., the bound given in (1) does not hold in this case, otherwise our bound would be innocuous, since

$$\sqrt{1 - \frac{\lambda_1}{\lambda_n}} \geq \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}.$$

Let us see an example.

Consider the quadratic function given in (7), with $Q = \begin{pmatrix} 1 & 0 \\ 0 & \lambda \end{pmatrix}$, $\lambda > 1$ and the sequence (x^k) generated by the Cauchy algorithm 3.1.

Let $x^k = (s \quad t)^T$ be the current point. So, $d^k = -\nabla f(x^k) = -(s \quad t\lambda)^T$ and $\alpha_k = (s^2 + \lambda^2 t^2)/(s^2 + \lambda^3 t^2)$. It can be easily seen that

$$x^{k+1} = x^k + \alpha_k d^k = \frac{(\lambda - 1)st}{s^2 + \lambda^3 t^2} \begin{pmatrix} \lambda^2 t \\ -s \end{pmatrix}. \tag{11}$$

Therefore (we assume $st \neq 0$, since otherwise $x^{k+1} = x^*$),

$$\frac{x_1^{k+1}}{x_2^{k+1}} = -\lambda^2 \frac{x_2^k}{x_1^k}. \tag{12}$$

If we let $m = x_1^0/x_2^0$, we can verify by induction and relation (12) that for all $k \in \mathbb{N}$,

$$\frac{x_1^{2k}}{x_2^{2k}} = m \qquad \text{and} \qquad \frac{x_1^{2k+1}}{x_2^{2k+1}} = -\frac{\lambda^2}{m}.$$

Hence, the sequence (x^k) alternates between the two straight lines $x_2 = x_1/m$ and $x_2 = -mx_1/\lambda^2$, as shown in Figure 3.

As we know that the Cauchy directions are orthogonal, that is, $(d^k)^T d^{k+1} = 0$, we can conclude that the triangles $\triangle x^{k+1} O x^k$ and $\triangle x^{k+3} O x^{k+2}$ are similar. Thus

$$\frac{\|x^{k+1}\|_2}{\|x^k\|_2} = \frac{\|x^{k+3}\|_2}{\|x^{k+2}\|_2}, \tag{13}$$

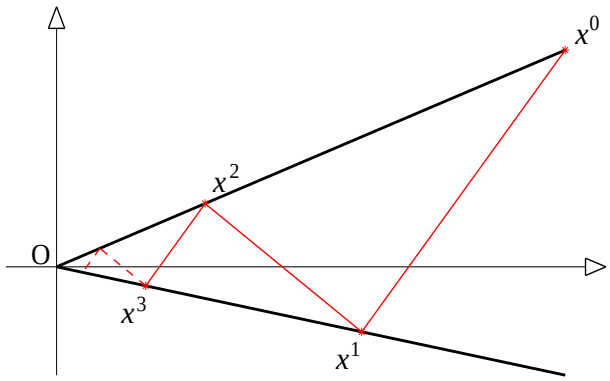

Figure 3: The path of (x^k) along two lines.

for all $k \in \mathbb{N}$. Furthermore, we have from (11) that

$$\frac{\|x^{k+1}\|_2}{\|x^k\|_2} = \frac{(\lambda-1)|st|}{s^2+\lambda^3 t^2}\sqrt{\frac{s^2+\lambda^4 t^2}{s^2+t^2}}, \tag{14}$$

which will be greater than $(\lambda-1)/(\lambda+1)$. For example, if $\lambda = 50$ and $x^0 = (20\ \ 1)^T$, we have, using (13) and (14), that

$$\frac{\|x^{2k+1}\|_2}{\|x^{2k}\|_2} = \frac{\|x^1\|_2}{\|x^0\|_2} > 0.9753, \tag{15}$$

for all $k \in \mathbb{N}$, whereas $(\lambda-1)/(\lambda+1) < 0.9608$. Note that $\sqrt{1-1/\lambda} \approx 0.9899$.

There is still a question that remains open for us. Could the bound $\gamma = \sqrt{1-\lambda_1/\lambda_n}$, given in Theorem 3.1, be reached? If not, what is the best bound? (Indeed, we have gotten strong evidences that there is a lower bound than γ).

Acknowledgements. The authors thank Sandra A. Santos and Clóvis C. Gonzaga for their valuable comments and suggestions which very much improved this paper.

References

1. J. Barzilai and J. M. Borwein. Two-point step size gradient methods. *IMA J. Numer. Anal.*, 8:141–148, 1988.

2. M. S. Bazaraa, H. D. Sherali, and C. M. Shetty. *Nonlinear Programming Theory and Algorithms*. John Wiley, New York, 2nd edition, 1993.

3. D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, Massachusetts, 1995.

4. 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.*, 151:862–880, 1999.

5. E. G. Birgin, J. M. Martínez, and M. Raydan. Nonmonotone spectral projected gradient methods on convex sets. *SIAM J. Optimization*, 10:1196–1211, 2000.

6. R. Burachik, L. Drummond, A. Iusem, and B. Svaiter. Full convergence of the steepest descent method with inexact line searches. *Optimization*, 32:137–146, 1995.

7. Y. H. Dai. Alternate step gradient method. Technical 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, China, 2001.

8. Y. H. Dai and R. Fletcher. On the asymptotic behaviour of some new gradient methods. Technical 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, China, 2003.

9. Y. H. Dai and L. Z. Liao. R-linear convergence of the Barzilai and Borwein gradient method. *IMA J. Numer. Anal.*, 26:1–10, 2002.

10. Y. H. Dai and X. Q. Yang. A new gradient method with an optimal stepsize property. Technical report, Hong Kong Polytechnic University, Hong Kong, 2002.

11. C. C. Gonzaga. Two facts on the convergence of the Cauchy algorithm. *Journal of Optimization Theory and Applications*, 107(3):591–600, 2000.

12. K. C. Kiwiel and K. G. Murty. Convergence of the steepest descent method for minimizing quasi convex functions. *Journal of Optimization Theory and Applications*, 89:221–226, 1996.

13. D. G. Luenberger. *Linear and Nonlinear Programming.* Addison - Wesley Publishing Company, New York, 1986.

14. J. M. Martínez and S. A. Santos. Métodos computacionais de otimização. $20.^0$ Colóquio Brasileiro de Matemática - IMPA, July 1995. In Portuguese.

15. M. Raydan. The Barzilai and Borwein gradient method for large scale unconstrained minimization problem. *SIAM Journal on Optimization*, 7:26–33, 1997.