Electronic Transactions on Numerical Analysis.
Volume 16, pp. 186-193, 2003.

ISSN 1068-9613.

ETNA
Kent State University
etna@mcs.kent.edu

GRADIENT METHOD WITH DYNAMICAL RETARDS FOR LARGE-SCALE OPTIMIZATION PROBLEMS*

FRANCISCO LUENGO† AND MARCOS RAYDAN‡

Abstract. We consider a generalization of the gradient method with retards for the solution of large-scale unconstrained optimization problems. Recently, the gradient method with retards was introduced to find global minimizers of large-scale quadratic functions. The most interesting feature of this method is that it does not involve a decrease in the objective function, which allows fast local convergence. On the other hand, nonmonotone globalization strategies, that preserve local behavior for the nonquadratic case, have proved to be very effective when associated with low storage methods. In this work, the gradient method with retards is generalized and combined in a dynamical way with nonmonotone globalization strategies to obtain a new method for minimizing nonquadratic functions, that can deal efficiently with large problems. Encouraging numerical experiments on well-known test problems are presented.

Key words. spectral gradient method, nonmonotone line search, Barzilai-Borwein method, Polak-Ribière method, Rayleigh quotient.

AMS subject classifications. 49M07, 49M10, 65K.

1. Introduction.

We consider the unconstrained minimization problem

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

where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and its gradient is available. We are interested in the large-scale case for which the Hessian of f is either not available or requires a prohibitive amount of storage and computational cost.

Spectral gradient methods, introduced by Barzilai and Borwein [1] and analyzed by Raydan [12], have a number of interesting features that make them attractive for the numerical solution of (1.1). The most important features of these methods are that only gradient directions are used, that the memory requirements are minimal, and that they do not involve a decrease in the objective function, which allows fast local convergence. They have been applied succesfully to find local minimizers of large scale real problems ([2, 3, 4, 5, 9, 14]).

In a recent paper, Friedlander, Martínez, Molina and Raydan [7] extended the spectral gradient methods and introduced the Gradient Method with Retards (GMR) to find the unique global minimizer of quadratic functions of the form

$$f(x) = \frac{1}{2}x^t Ax - b^t x \ , \tag{1.2}$$

where $A \in \mathbb{R}^{n\times n}$ is large, sparse, symmetric and positive definite (SPD). The GMR is in fact a large class of methods that can be written as

$$x_{k+1} = x_k - \frac{1}{\alpha_{\nu(k)}} g_k \ , \tag{1.3}$$

where g_k is the gradient vector of f evaluated at x_k, and

$$\alpha_{\nu(k)} = \frac{g^t_{\nu(k)} A^{\rho(k)} g_{\nu(k)}}{g^t_{\nu(k)} A^{(\rho(k)-1)} g_{\nu(k)}} \ . \tag{1.4}$$

*Received January 9, 2002. Accepted for publication June 27, 2003. Recommended by D. Sorensen.

†Dpto. de Computación, Facultad de Ciencias, Universidad del Zulia, Ap. 527, Maracaibo, Venezuela. E-mail: fluengo@cantv.net.

‡Dpto. de Computación, Facultad de Ciencias, Universidad Central de Venezuela, Ap. 47002, Caracas 1041-A, Venezuela. E-mail: mraydan@reacciun.ve. This research was partially supported by UCV-PROJECT 97-003769.

ETNA
Kent State University
etna@mcs.kent.edu

In (1.4), $\nu(k)$ and $\rho(k)$ are arbitrarily chosen in the sets

$$\{k, k-1, \ldots, max\{0, k-\bar{m}\}\}, \tag{1.5}$$

and

$$\{q_1, q_2, \ldots, q_{\bar{m}}\}, \tag{1.6}$$

respectively, where $\bar{m}$ is a given positive integer, and q_i is a given positive integer for $i = 1, 2, \ldots, \bar{m}$. For instance, when $\rho(k) = 1$ for all k and $\nu(k) = k$, then the GMR reduces to the classical steepest descent method or Cauchy method. If $\rho(k) = 1$ for all k and $\nu(k) = k-1$, then the GMR becomes the spectral gradient method. Extensive numerical results discussed in [7] indicate that some members of the GMR family clearly outperform the spectral gradient method, requiring less computational work.

For the nonquadratic case, the GMR needs to be incorporated in a globalization strategy. Since the method does not enforce decrease in the objective function, a nonmonotone line search strategy will be used. In particular, the nonmonotone line search technique introduced by Grippo, Lampariello and Lucidi [10] has proved to be very effective for large-scale optimization problems. This line search essentially enforces the following condition

$$f(x_{k+1}) \leq \max_{0 \leq j \leq M} f(x_{k-j}) + \gamma g_k^t(x_{k+1} - x_k), \tag{1.7}$$

where M is a nonnegative integer and γ is a small positive number.

The new algorithm combines and extends the following results: the globalization of the spectral gradient method by Raydan in [13] that is based on (1.7), and the gradient method with retards by Friedlander, Martínez, Molina and Raydan [7]. In particular, a special dynamical retard that takes advantage of the approximating property of eigenvectors is developed and included in the algorithm.

The rest of the paper is divided into sections as follows. In Section 2 we present a general nonmonotone gradient algorithm for which we can establish classical convergence results. In Section 3 we present our dynamical version of the global gradient method with retards for the minimization of nonquadratic functions. In Section 4 we discuss implementation details and show numerical results on some classical test functions. Finally, in Section 5 we present some final remarks.

2. Nonmonotone gradient methods. For the minimization of nonquadratic functions, nonmonotone methods like the spectral gradient method, need to be incorporated with a globalization strategy. Raydan [13] proposed a globalization scheme for the spectral gradient algorithm that fits nicely with the nonmonotone behavior of this family of methods. Roughly speaking, the main idea is to accept the step if it satisfies a weak condition of the form given by (1.7). When $M > 0$ this condition allows the objective function to increase at some iterations and still guarantees global convergence, as we discuss later. First, in this section, we would like to present a general nonmonotone gradient algorithm for which we can establish classical convergence results.

Algorithm 2.1: Global Nonmonotone Gradient Method
Given x_0, $0 < \epsilon < 1$, $\alpha_0 \in [\epsilon, 1/\epsilon]$, integer $M > 0$, $\gamma \in (0,1)$, $0 < \sigma_1 < \sigma_2 < 1$.
Set $k = 0$.
while $(\|g_k\| > tolerance)$ **do**

set $\lambda = 1/\alpha_k$

while $(f(x_k - \lambda g_k) > \max_{0\leq j\leq min(k,M)}\{f_{k-j}\} - \gamma\lambda g_k^t g_k)$ **do**

choose $\sigma \in [\sigma_1, \sigma_2]$, and **set** $\lambda = \sigma\lambda$.

endwhile

set $\lambda_k = \lambda$, and $x_{k+1} = x_k - \lambda_k g_k$.

set $g_{k+1} = \bigtriangledown f(x_{k+1})$

choose $\alpha_{k+1} \in [\epsilon, 1/\epsilon]$, and **set** $k = k + 1$

endwhile

We will describe various possibilities for the parameter *tolerance* later. The choice of α_{k+1} in Algorithm 2.1 is quite general. Specific choices of steplength related to the gradient method with retards that produce fast convegence will be discussed in the next section. However, even at this level of generality we can establish a global convergence result.

THEOREM 2.1. *Assume that* $\Omega = \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ *is a bounded set. If* f *is continuously differentiable on an open set that contains* Ω*, then Algorithm 2.1 is well defined and any accumulation point of the sequence* $\{x_k\}$ *that it generates is a stationary point.*

Proof. If x_k is not a stationary point, then $g_k^t g_k > 0$. Since $\gamma < 1$, then for sufficiently small values of λ the following condition holds:

$$f(x_k - \lambda g_k) \leq \max_{0\leq j\leq \min(k,M)} (f_{k-j}) - \gamma\lambda g_k^t g_k.$$

Hence, a stepsize satisfying the nonmonotone line search will be found after a finite number of trials, and Algorithm 2.1 is well defined.

For the second part of the proof notice that, choosing $G_k = I$ for all k and $\delta \in [\epsilon, 1/\epsilon]$, the Global Preconditioned Spectral Gradient algorithm (Basic version) in [11, pp. 243-244] reduces to Algorithm 2.1. Therefore, the result follows from Theorem 2.1 in [11, p. 244]. ◻

In a recent work, Dai [6] presents additional convergence results for nonmonotone line search techniques under additional assumptions. In particular, he establishes an R-linear rate of convergence result for uniformily convex functions.

3. Global gradient method with retards. Motivated by the GMR method introduced in [7] for convex quadratics, we can extend (1.3) and present methods of the following type

$$x_{k+1} = x_k - (1/\alpha_{\nu(k)})g_k, \tag{3.1}$$

where

$$\alpha_{\nu(k)} = \frac{s_{\nu(k)}^t y_{\nu(k)}}{s_{\nu(k)}^t s_{\nu(k)}}, \tag{3.2}$$

$s_{\nu(k)} = x_{\nu(k)+1} - x_{\nu(k)}$, $y_{\nu(k)} = g_{\nu(k)+1} - g_{\nu(k)}$ and $\nu(k)$ is arbitrary chosen in the set $\{0, 1, \ldots, k\}$.

For the convex quadratic case, $\nu(k)$ is chosen in the set (1.5) and the GMR method converges to global minimizers. A direct extension of the GMR for quadratics, combined with the nonmonotone line search, produces our first version of the global GMR that can be

ETNA
Kent State University
etna@mcs.kent.edu

obtained by substituting the choice of α_{k+1} in Algorithm 2.1 by the following step:

Choose $\alpha_{k+1} \in \{(s_0^t y_0)/(s_0^t s_0), \ldots, (s_k^t y_k)/(s_k^t s_k)\}$, and **set** $k = k + 1$.

At every iteration we need to fit the scalar $(s_k^t y_k)/(s_k^t s_k)$ in the interval $[\epsilon, 1/\epsilon]$, as follows: if $\alpha_k \leq \epsilon$ or $\alpha_k \geq \frac{1}{\epsilon}$, then $\alpha_k = \delta$, for a given positive $\delta \in [\epsilon, 1/\epsilon]$.

Based on the theory discussed in the previous section, global convergence is established for this particular case. However, it is still quite general for practical purposes. Indeed, extensive numerical experimentation in [7] indicate that the use of *very old* retards deteriorates the speed of convergence. Moreover, they observed that a retard that oscillates between 1 and 2 is ideal for fast convergence. On the other hand, it would be interesting to detect the presence of approximated eigenvectors as descent directions. We have seen, and it has been discussed by Glunt et al. [9], that in the spectral gradient case the approximation of eigenvectors frequently happens. In fact, if the gradient direction is an eigenvector of $\nabla^2 f(x_k)$ then the best option is by all means the Cauchy choice of steplength, i.e., no retard, since in the quadratic case it would imply termination at the solution in the next iteration. This motivates us to include the following strategy for choosing the steplength.

- If x_k is "close to the solution" and

$$cos_k \equiv \frac{|g_k^t y_k|}{\|g_k\| \, \|y_k\|} \approx 1,$$

  then $\alpha_k = s_k^t y_k / s_k^t s_k$ (steepest descent).
  The motivation for this choice is that if $cos_k \approx 1$ then g_k is a good approximation to an eigenvector of $\nabla^2 f(x_k)$ and the steepest descent (Cauchy) choice is an excellent choice. Indeed, cos_k is the cosine of the angle between g_k and the average Hessian matrix $\int_0^1 \nabla^2 f(x_{k-1} + t s_{k-1}) dt$ times g_k. Therefore, if $cos_k = 1$, then g_k is an eigenvector of the average Hessian matrix. Close to the solution, the average Hessian matrix closely approximates the Hessian of f at the solution, and the Cauchy choice of steplength would closely approximate the minimizer.
- If $cos_{k-1} \geq cos_k$, then (double retard), i.e.,

$$\alpha_{k+1} = s_{k-1}^t y_{k-1} / s_{k-1}^t s_{k-1}.$$

  The motivation here is that the bigger the cosine the closer the gradient to an eigenvector. Moreover, α_{k+1} would be closer to an eigenvalue, and we obtain longer steps, which is suitable once we are close to the solution and the norm of the gradient is close to zero.
- Otherwise $\alpha_{k+1} = s_k^t y_k / s_k^t s_k$ (spectral choice).

All the previous remarks lead us to the following algorithm.

Algorithm 3.1: GGMR
Given x_0, $0 < \epsilon < 1$, $\alpha_0 \in [\epsilon, 1/\epsilon]$, integer $M > 0$, $\gamma \in (0, 1)$, $0.9 < L < 1$,
$\delta > 0$, $0 < \sigma_1 < \sigma_2 < 1$.
Set $k = 0$, $\alpha_{new} = \alpha_0$, $oldc = 0$, $newc = 0$.

while $(\|g_k\| > tolerance)$ **do**
  set $\lambda = 1/\alpha_k$

```
/* Nonmonotone Line Search */
```
while $(f(x_k - \lambda g_k) > \max_{0 \le j \le min(k,M)}\{f_{k-j}\} - \gamma\lambda g_k^t g_k)$ **do**
  choose $\sigma \in [\sigma_1, \sigma_2]$, and **set** $\lambda = \sigma\lambda$.
endwhile

set $\lambda_k = \lambda$, and $\bar{x} = x_k - \lambda_k g_k$
set $\bar{g} = \bigtriangledown f(\bar{x})$, and $y = \bar{g} - g_k$

```
/* Updating the cosines */
```
if ("close to the solution") **then**
  set $oldc = newc$, and $newc = |g_k^t y|/(\|g_k\|\|y\|)$
endif

set $\alpha_{new} = -(g_k^t y)/(\lambda_k g_k^t g_k)$

if $(\alpha_{new} \le \epsilon$ or $\alpha_{new} \ge 1/\epsilon)$ **then**
  set $\alpha_{new} = \delta \in [\epsilon, 1/\epsilon]$
endif

```
/* Cauchy choice */
```
if ("close to the solution" and $newc \ge L$) **then**
  set $\alpha_k = \alpha_{new}$
else
  ```
  /* Retarded choice */
  ```
  if ("close to the solution" and $oldc \ge newc$) **then**
    set $\alpha_{k+1} = \alpha_k$
  else
    set $\alpha_{k+1} = \alpha_{new}$
  endif

  set $g_{k+1} = \bigtriangledown f(x_{k+1})$
  set $x_{k+1} = \bar{x}$, and $k = k + 1$
endif
endwhile

We would like to close this section with the most relevant characteristics of Algorithm 3.1:

1. Requires $4n$ storage locations.
2. Every iteration requires $O(n)$ flops and one gradient evaluation, unless the Cauchy choice is used. In that iteration, the gradient is evaluated twice.
3. $f(x_k)$ is *not* monotonically decreasing.
4. Since $s_k = -\lambda_k g_k$, then the definition of α_{new} in the algorithm is equivalent to the one given in (3.2). The advantage of this equivalent expression is that it avoids the storage of the vector s_k.

4. Numerical Results. We compare the Global Gradient Method with Retards (GGMR) with the Global Spectral Gradient (GSG) [13], and the Polak-Ribiere implementation (PR^+) of Gilbert and Nocedal [8], on some classical test functions listed on Table 4.1. The PR^+ code requires $4n$ storage locations and it is, to the best of our knowledge, the most effective implementation of the conjugate gradient method for nonquadratic functions. The GSG can be viewed as a particular case of Algorithm 3.1, ignoring the Cauchy choice and the retarded choice.

All experiments were run on a Pentium III at 750Mhz, 128MRAM, and double precision Fortran 77. We used the following stopping criterion:

$$\|g_k\|_2 \leq 10^{-8}(1 + |f(x_k)|).$$

For our numerical experiments, "close to the solution" means

$$\|g_k\|_2 \leq 10^{-2}(1 + |f(x_k)|),$$

and the following parameters were used: $\gamma = 10^{-4}$, $M = 10$, $\sigma_1 = 0.1$, $\sigma_2 = 0.5$, $\epsilon = 10^{-10}$, $\alpha_0 = 1$, and $L = 0.95$. The parameter δ is chosen as follows:

$$\delta = \max\{1, \min\{1/\epsilon, \|g_k\|\}\}.$$

For descriptions of the test functions and the starting points, see [8].

Problem	Name
1	Brown almost linear
2	Broyden tridiagonal
3	Extended ENGLV1
4	Extended Freudenstein and Roth
5	Generalized Rosenbrock
6	Oren's power
7	Penalty 1
8	Extended Powell singular
9	Extended Rosenbrock
10	Tridiagonal 1
11	Trigonometric
12	Variably dimensioned
13	Wrong extended Wood
14	Strictly convex 1
15	Strictly convex 2
16	Extended Box 3-D
17	Extended Biggs EXP6

TABLE 4.1
Classical test functions

The numerical results are shown in Tables 4.2 and 4.3. We report the function and the dimension (f/n), the number of iterations required for convergence (It), the number of gradient evaluations (g), the number of iterations for which the retard choice was used (Rt), and the number of iterations for which the Cauchy or steepest descent choice was used (SD). For GSG and GGMR the number of iterations and gradient evaluations are equal, and so it is reported under the label (It/g). The asterisk (*) that appears under the multicolumn PR^+ means that the method could not find a local solution after 5000 iterations. The results of Tables 4.2 and 4.3 are summarized in Table 4.4. We report, in Table 4.4, the number of problems for which each method was a winner in number of iterations, and number of gradient evaluations.

We observe that GSG and GGMR are both very robust for finding local minimizers of large-scale nonquadratic functions. GSG failed to converge only for function 11 and $n =$

ETNA
Kent State University
etna@mcs.kent.edu

	PR^+		**GSG**	**GGMR**		
f/n	It	g	It/g	It/g	Rt	SD
1/1000	*	*	4	3	0	1
1/10000	*	*	8	5	0	3
2/1000	43	90	47	45	26	1
2/5000	43	91	107	58	25	0
3/1000	19	57	33	28	14	4
3/10000	*	*	28	26	15	3
4/1000	13	43	4823	263	103	4
4/10000	*	*	133	132	2	4
5/100	275	574	1711	1025	544	2
5/500	1089	2202	4603	4037	2097	2
6/1000	130	268	376	333	198	7
6/10000	23	89	63	46	0	22
7/1000	*	*	63	31	12	26
7/10000	*	*	73	43	28	35
8/1000	152	367	2172	468	225	44
8/10000	*	*	3380	482	250	42
9/1000	25	74	101	35	25	9
9/10000	*	*	69	60	7	5

TABLE 4.2
PR^+, GSG, and GGMR on classical test functions

	PR^+		**GSG**	**GGMR**		
f/n	It	g	It/g	It/g	Rt	SD
10/1000	332	667	131	630	337	2
10/10000	239	484	131	133	70	0
11/1000	*	*	111	111	57	1
11/10000	*	*	*	67	30	0
12/100	*	*	38	22	7	21
12/500	*	*	55	28	12	2
13/1000	35	80	79	66	33	3
13/10000	*	*	61	70	40	3
14/1000	5	18	8	6	0	3
14/10000	4	15	8	6	0	3
15/1000	107	218	237	162	83	3
15/10000	14	45	19	19	5	1
16/300	51	155	85	44	20	10
16/3000	140	327	898	50	27	12
17/600	131	319	2881	678	346	32
17/1200	67	181	2313	692	179	14

TABLE 4.3
PR^+, GSG, and GGMR on classical test functions

10000, and GGMR never failed to converge. They both outperform PR^+ in number of gradient evaluations, except for problems with a very ill-conditioned Hessian at the solution. For some of these problems, GGMR is still competitive. However, if the Hessian is singular at the solution as in functions 8 and 17, then PR^+ clearly outperforms GSG and GGMR.

Method	IT	g
PR^+	17	7
GSG	3	3
GGMR	13	22

TABLE 4.4
Number of winners for each method

On the other hand, PR^+ outperforms GSG and GGMR in number of iterations, except for some problems where PR^+ failed to converge before 10000 iterations.

We also observe that GGMR outperforms GSG in most cases (28 out of 36). In some of those cases, the difference between them is remarkable.

Finally, we would like to comment that for PR^+ a line search is required at every iteration. Whereas for GSG and GGMR a line search is needed at very few iterations. For very ill-conditioned problems they both require approximately a line search for every 5 iterations, which implies a significant reduction in CPU time.

5. Final remarks. The recently developed GMR method can be globalized with a nonmonotone line search to produce fast convergent gradient methods for the minimization of nonquadratic functions. In particular, the feature of approximating eigenvalues and eigenvectors help to accelerate the convergence of gradient related methods. Our numerical results indicate that the globalized GMR new method represents a fast and robust option for unconstrained optimization, that requires few line search.

REFERENCES

[1] J. BARZILAI AND J. BORWEIN, *Two point step size gradient methods*, IMA J. Numer. Anal., 8 (1988), pp. 141–148.
[2] E. BIRGIN, I. CHAMBOULEYRON, AND J. MARTINEZ, *Estimation of the optical constants and the thickness of thin films using unconstrained optimization*, J. Comput. Phys., 151 (1999), pp. 862–880.
[3] E. BIRGIN AND Y. EVTUSHENKO, *Automatic differentiation and spectral projected gradient methods for optimal control problems*, Optim. Methods and Softw., 10 (1998), pp. 125–146.
[4] Z. CASTILLO, D. CORES, AND M. RAYDAN, *Low cost optimization techniques for solving the nonlinear seismic reflection tomography problem*, Optim. Eng., 1 (2000), pp. 155–169.
[5] D. CORES, G. FUNG, AND R. MICHELENA, *A fast and global two point low storage optimization technique for tracing rays in 2D and 3D isotropic media*, J. Appl. Geophys., 45 (2000), pp. 273–287.
[6] Y. H. DAI, *On nonmonotone line search*, Tech. Rep. ICM-01-07, Laboratory of Scientific and Engineering Computing, Chinese Academy of Sciences, Beijing, China, 2001.
[7] A. FRIEDLANDER, J. MARTINEZ, B. MOLINA, AND M. RAYDAN, *Gradient method with retards and generalizations*, SIAM J. Numer. Anal., 36 (1999), pp. 275–289.
[8] J. GILBERT AND J. NOCEDAL, *Global convergence properties of conjugate gradient methods for optimization*, SIAM J. Optim., 2 (1992), pp. 21–42.
[9] W. GLUNT, T. HAYDEN, AND M. RAYDAN, *Molecular conformations from distance matrices*, J. Comput. Chem., 14 (1993), pp. 114–120.
[10] 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.
[11] F. LUENGO, M. RAYDAN, W. GLUNT, AND T. HAYDEN, *Preconditioned spectral gradient method*, Numer. Algorithms, 30 (2002), pp. 241–258.
[12] M. RAYDAN, *On the Barzilai and Borwein choice of steplength for the gradient method*, IMA J. Numer. Anal., 13 (1993), pp. 321–326.
[13] ———, *The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem*, SIAM J. Optim., 7 (1997), pp. 26–33.
[14] C. WELLS, W. GLUNT, AND T. HAYDEN, *Searching conformational space with the spectral distance geometry algorithm*, J. Molecular Structure (Theochem), 308 (1994), pp. 263–271.