International Journal of Applied Mathematics

Volume 33 No. 5 2020, 853-866
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
doi: http://dx.doi.org/10.12732/ijam.v33i5.8

NEW TWO-STEP CONJUGATE GRADIENT METHOD FOR UNCONSTRAINED OPTIMIZATION

Issam A.R. Moghrabi

Gulf University for Science and Technology
College of Business Administration
Mishref - 1150, KUWAIT

Abstract: Two-step methods are secant-like techniques of the quasi-Newton type that, unlike the classical methods, construct nonlinear alternatives to the quantities used in the so-called Secant equation. Two-step methods instead incorporate data available from the two most recent iterations and thus create an alternative to the Secant equation with the intention of creating better Hessian approximations that induce faster convergence to the minimizer of the objective function. Such methods, based on reported numerical results published in several research papers related to the subject, have introduced substantial savings in both iteration and function evaluation counts. Encouraged by the successful performance of the methods, we explore in this paper employing them in developing a new Conjugate Gradient (CG) algorithm. CG methods gain popularity on big problems and in situations when memory resources are scarce. The numerical experimentations on the new methods are encouraging and open venue for further investigation of such techniques to explore their merits in a multitude of applications.

AMS Subject Classification: 65K10
Key Words: weighted Conjugate Gradient methods; quasi-Newton Methods; multi-step methods; unconstrained optimization

Received: June 30, 2020

1. Introduction

This work addresses problems of the form:

$$\text{minmize } f(x),\ x \in R^n,\ \text{where } f : R^n \to R.$$

Conjugate Gradient methods (CG) are a class of methods for solving large unconstrained optimization problems. Their storage requirements are rather minimal compared to other methods as they do not store any matrices. While such methods converge to the minimum in at most n iterations on quadratic functions for accurate line searches, they are also employed to optimize non-quadratic functions where line searches are not usually exact. In the context of minimizing non-quadratic functions, the methods need to be restarted when certain criteria are met (see [22]). Such methods have been the subject of many research papers in which variations to the original method of Fletcher and Reeves [9] have been examined (see, for example, [2,5,13,14]).

To minimize f, the sequence of iterates generated is given by

$$x_{i+1} = x_i + \alpha_i d_i, \tag{1}$$

where α_i is a positive scalar and d_i is a CG search direction. The search direction is computed using

$$d_i = \begin{cases} -g_i, & \text{for } i = 0, \\ -g_i + \beta_i d_{i-1}, & \text{for } i \geq 1, \end{cases} \tag{2}$$

for some chosen scalar β_i and where g_i denotes the gradient of the function f at x_i. The search direction d_i is normally required to satisfy

$$d_i^T g_i < 0,$$

to ensure it is a downhill direction. In order to ensure global convergence, d_i may be required to satisfy the sufficient downhill condition

$$d_i^T g_i \leq -\xi \|g_i\|^2,$$

for some positive constant ξ.

Different formulae for β_i lead to different CG algorithms. Some published choices are

$$\beta_i^{FR} = \frac{\|g_i\|^2}{\|g_{i-1}\|^2}, \quad \beta_i^{PRP} = \frac{g_i^T (g_i - g_{i-1})}{\|g_{i-1}\|^2},$$

$$\beta_i^{HS} = \frac{g_i^T(g_i - g_{i-1})}{d_{i-1}^T(g_i - g_{i-1})}, \quad \beta_i^{LS} = \frac{g_i^T(g_i - g_{i-1})}{d_{i-1}^T g_{i-1}},$$

$$\beta_i^{DY} = \frac{g_i^T g_i}{d_{i-1}^T(g_i - g_{i-1})},$$

in addition to many other suggestions (see, for example, [13], [16], [20], [21], [22], [23]). The above formulae are, respectively, due to Fletcher-Reeves [9], Hestenes-Stiefel [15], Polyak [22], Polak–Ribiére [21], Liu-Storey [16] and Dai–Yuan [5].

This paper derives a new CG algorithm that is based on the multi-step approach while avoiding the storage of any update matrix. Anderi [2] applies updates to the identity matrix to build the conditioning matrix. Ford et al. [13] develop multi-step CG methods that do not involve any weighting matrices. Our derivation follows rather a different approach. The next section introduces the idea of multi-step and the specific two-step methods. The derivation of the new method is done in Section 3. Finally, the numerical results are presented followed by the conclusions.

2. Multi-step Quasi-Newton methods

Quasi-Newton methods update a step-wise approximation to the Hessian of the objective function $n \times n$ matrix that exploits the latest computed data, [3]. Given B_i, the most recent approximation to the Hessian, the next Hessian approximation, B_{i+1} is classically updated to satisfy the classical secant relation:

$$B_{i+1}s_i = y_i, \tag{3}$$

where

$$s_i = x_{i+1} - x_i,$$

and

$$y_i = g_{i+1} - g_i.$$

The BFGS formula [3,10,11] is the most popular update formula that s that satisfies (3) especially that it is the most successful update that performs best so far with inaccurate line search algorithms [3,13,14]. This rank-two update that approximates the actual Hessian is given by

$$B_{i+1} = B_i - \frac{B_i s_i s_i^T B_i}{s_i^T B_i s_i} + \frac{y_i y_i^T}{s_i^T y_i}.$$

A straight line L is usually used in the standard secant equation to find a new point x_{i+1}, given the previous point x_i, while higher order polynomials are employed in the construction of the multi-step techniques [11,12,13].

Let $\{x(\tau)\}$, denoted by X, define a differentiable path in R^n, for $\tau \in R$. The polynomial $x(\tau)$ is chosen to satisfy

$$x(\tau_j^{(i-1)}) = x_{i-m+j}, \text{ for } j = 0, 1, \ldots, m,$$

for some non-overlapping values $\left\{\tau_j^{(i-1)}\right\}_{j=0}^{m}$. The corresponding gradient points are defined by a similar polynomial $z(\tau)$ satisfying

$$z(\tau_j^{(i-1)}) = g_{i-m+j}, \text{ for } j = 0, 1, \ldots, m.$$

If the Chain rule is applied to the gradient vector $z(x(\tau)) \approx g(x(\tau))$, we obtain

$$\frac{dg}{d\tau} = G(x(\tau))\frac{dx}{d\tau}. \tag{4}$$

This indicates that, for any point on the path X, the Hessian matrix G satisfies (4) for any chosen value of the parameter τ. In particular, for $\tau = \tau_c$, where $\tau_c \in \mathcal{R}$ the following is obtained

$$\frac{dg}{d\tau}_{\tau=\tau_c} = G(x(\tau))\frac{dx}{d\tau}_{\tau=\tau_c}.$$

To derive a relation similar to the secant equation the new Hessian approximation must satisfy at the new point x_{i+1}, some value for the τ-parameter, namely τ_m, that corresponds to the latest iterate, is selected such that

$$g'(\tau_m) = B_{i+1}x'(\tau_m),$$

or

$$w_i = B_{i+1}r_i, \tag{5}$$

where the vectors r_i and w_i are constructed using the latest m step vectors $\{s_k\}_{k=i-m+1}^{i}$ and the respective m gradient difference vectors $\{y_k\}_{k=i-m+1}^{i}$, as follows (where $L_k(\tau)$ denote the standard Lagrange polynomials)

$$r_i = \sum_{j=0}^{m-1} s_{i-j} \left\{ \sum_{k=m-j}^{m} L_k^{'}(\tau_m) \right\}$$

and

$$w_i = \sum_{j=0}^{m-1} y_{i-j} \left\{ \sum_{k=m-j}^{m} L_k^{'}(\tau_m) \right\}$$

for

$$L_k^{'}(\tau_m) = (\tau_k - \tau_m)^{-1} \left[\frac{\tau_m - \tau_j}{\tau_k - \tau_j} \right], \quad k < m,$$

and

$$L_m^{'}(\tau_m) = \sum_{j=0}^{m-1} (\tau_m - \tau_j)^{-1}.$$

Several choices for the parameters $\{\tau_k\}_{k=0}^{m}$ have been considered in Ford and Moghrabi [13]. Such choices are sensitively influential to the structure of the interpolating curve. In this paper, the choice based on the so-called Accumulative Approach is elected due to its numerical merits [18,19].

The choices taken into consideration for the values $\{\tau_k\}_{k=0}^{m}$ are based on a metric expressed using

$$\emptyset_M(z_1, z_2) = \left[(z_1 - z_2)^T M (z_1 - z_2)\right]^{1/2},$$

where the matrix M is some symmetric positive-definite matrix.

The accumulative approach considered here focuses on one of the points, say x_j, and chooses its corresponding parameter τ_j to be zero. Then, any τ_k, that corresponds to the point $x_{i-m+k+1}$, for all k except for $k = j$, is obtained by the accumulation of the distance (calculated using $\emptyset_M$) between every adjacent pair of iterates in the sequence from $x_{i-m+j+1}$ to $x_{i-m+k+1}$. Therefore, for $k = 0,1,\ldots,m$, any parameter τ_k is calculated as

$$\begin{aligned} \tau_k &= -\sum_{p=k+1}^{j} \emptyset_M(x_{i-m+p+1}, x_{i-m+p}), \quad k < j, \\ &= 0, \quad k = j, \\ &= -\sum_{p=j+1}^{k} \emptyset_M(x_{i-m+p+1}, x_{i-m+p}), \quad k > j. \end{aligned} \tag{6}$$

This technique yields τ-values that obey

$$\tau_k < \tau_{k+1}, \text{ for } k = 0,\ 1, \ldots, m-1.$$

such that no overlapping occurs for any pair of points.

Once the parameters $\{\tau_k\}$ are computed, the vectors $x\prime(\tau_m)$ and $g\prime(\tau_m)$ in (5) (or vectors r_i and w_i, respectively) are obtainable and hence the new Hessian estimated matrix B_{i+1} is computed so that (5) is satisfied.

It is worth noting that the values obtained for τ-parameters are sensitive to the choice of the matrix M in $\emptyset_M$. Different choices lead to different values and result in different numerical behaviour. Since Ford and Moghrabi [12,13] report that tests involving values of $m > 2$ have not yielded considerable numerical improvements, a thing that may be attributed to possibly the non-smoothness of the interpolant, the choice $m = 2$ is adopted in this work and such methods are termed as two-step methods. Two-step methods exploit data available from the two latest steps in updating the Hessian approximation.

Choices considered for the matrix M include $M = B_i$, $M = B_{i+1}$ and $M = I$, [12,18,19]. For the 2-step methods, the Hessian approximation is updated to generally satisfy:

$$H_{i+1}\left(y_i - \mu_{i-1}y_{i-1}\right) = s_i - \mu_{i-1}s_{i-1}, \tag{7}$$

or

$$w_i = B_{i+1}r_i),$$

where

$$\mu_{i-1} = \frac{\delta_{i-1}^2}{2\delta_{i-1} + 1}$$

and

$$\delta_{i-1} = \frac{\tau_2^{(i-1)} - \tau_1^{(i-1)}}{\tau_1^{(i-1)} - \tau_0^{(i-1)}}.$$

For our numerical tests, the choice $M = I$ is taken in (6) and hence

$$\tau_0 = -\left(\|s_i\|_2 + \|s_{i-1}\|_2\right),\ \tau_2 = 0,\ \text{and}\ \tau_1 = -\|s_i\|_2\ .$$

This yields

$$\delta = \frac{\|s_i\|}{\|s_{i-1}\|}. \tag{8}$$

More generally, (8) may be modified by plugging in a scaling parameter, $\gamma \geq 0$ that offers more control in this context since setting the scalar to zero provides a convenient switching to the standard secant one-step update. Therefore,

$$\delta = \gamma\frac{\|s_i\|}{\|s_{i-1}\|}. \tag{9}$$

The multi-step BFGS Hessian approximation update formula is given by

$$B_{i+1}^{MS} = B_i + \frac{w_i w_i^T}{w_i^T r_i} - \frac{B_i r_i r_i^T B_i}{r_i^T B_i r_i}. \tag{10}$$

3. A new multi-step CG method (MSCG)

In this work, the search direction under study has the form

$$d_i = -\sigma_i g_i + \beta_i s_{i-1}, \tag{11}$$

(see [2,3]), where σ_i may be taken to be some selected scalar or a carefully chosen positive definite matrix. To illustrate, if $\sigma_i = 1$, then (11)is equivalent to (2). If, on the other hand, σ_i is taken to be a matrix that approximates the inverse Hessian, then d_i combines both conjugate gradient and quasi-Newton direction vectors. In this work, we consider a variant of the latter case.

When the linear CG methods is applied to a quadratic function, the computed search vectors satisfy the following conjugacy condition

$$d_i^T A d_j = 0, \ \forall i \neq j, \tag{12}$$

where A is the constant Hessian of the quadratic objective function that is known to be positive definite. However, when the objective function is non-quadratic, a more reasonable choice than relation (12) may be taken such that (see [14])

$$d_i^T y_{i-1} = 0. \tag{13}$$

Perry [20] studied possible improvements to the CG methods by considering advantages of the quasi-Newton methods. His approach uses (3) and the fact that the quasi-Newton search direction is computed using $d_i = -H_i g_i$, then Perry replaced (13) with

$$d_i^T y_{i-1} = -g_i^T s_{i-1}. \tag{14}$$

Using (14), one gets

$$d_i^T y_{i-1} = -g_i^T \left(H_i y_{i-1}\right),$$

or (from (7))

$$d_i^T y_{i-1} = -g_i^T r_{i-1} - \mu_{i-1} g_i^T H_i y_{i-2}.$$

This yields

$$d_i^T w_{i-1} = -\epsilon g_i^T r_{i-1}, \tag{15}$$

for some $\varepsilon \geq 0$, used to impose conjugacy.

Upon substituting (compare to (11))

$$d_i = -\sigma_i g_i + \beta_i s_{i-1} \tag{16}$$

in (15), we obtain

$$-\sigma_i w_{i-1}^T g_i + \beta_i w_{i-1}^T s_{i-1} = -\epsilon g_i^T r_{i-1},$$

giving the following expression for β_i

$$\beta_i = \frac{g_i^T[\sigma_i w_{i-1} - \varepsilon r_{i-1}]}{s_{i-1}^T w_{i-1}}. \tag{17}$$

If $\varepsilon = 0$, then (17) reduces to the choice of β_i obtained in [13].

The choice $\varepsilon = 0$ is adopted here and σ_i is chosen to ensure that $\beta_i \geq 0$ for the method derived here. The search direction takes the form

$$d_{i+1} = -g_{i+1} + \beta_{i+1} s_i,$$

with β_{i+1} given by (17).

Global convergence of the Fletcher-Reeves method was proven by Al-Baali [1] on general functions with inexact line search. Dai and Yuan [5] derived a CG method derived using the secant condition for which global convergence was proven. In order to guarantee the convergence of the algorithm in [5], the step size α_i in (1) is deemed acceptable provided it satisfies the Wolfe conditions [28]:

$$f(x_i + \alpha_i d_i) - f(x_i) \leq \rho_1 \alpha_i d_i^T g_i, \tag{18}$$

$$g(x_i + \alpha_i d_i)^T d_i \geq \rho_2 d_i^T g_i, \tag{19}$$

where $0 < \rho_1 \leq \rho_2 < 1$.

The following lemma highlights the conditions that ensure the search direction is downhill.

Lemma 1. *The search direction given by(16) is a downhill direction.*

Proof. Given that $d_0 = -g_0$, it follows that $g_0^T d_0 = -\|g_0\|^2 \leq 0$.

As for later iterations, pre-multiplying (16) by $-g_{i+1}^T$ gives

$$-g_{i+1}^T d_i = -\sigma_i g_i + \beta_i s_{i-1}$$

which, assuming s_{i-1} is downhill, yields a downhill direction if σ_i is chosen to ensure that $\beta_i \geq 0$. □

4. Numerical computations

Our numerical results are benchmarked against those in Anderi's SCALCG [2]. The results reported in Table 2 are for different selections of the parameter γ in (9)in order to determine its effect on the numerical behavior of the method. The values that appear in Table 2 for γ correspond to 0, 0.5 and 1, respectively. Other values have been considered but they bear no significant changes. The numbers reported indicate iteration/function and gradient evaluations counts, respectively. The problems are primarily extracted from [1,2,5,13,14,20,25,26,27,29].

problem	**function name**	**size**
1	Extended Rosenbrock	100000
2	Extended Powell Singular	100000
3	Trigonometric	100000
4	Oren	10000
5	Cube	2
6	Extended Wood	4-100000
7	Beale	2
8	Helical Valley	3
9	Penalty I	2
10	Watson function	4
11	Variably dimensioned function	2-100000
12	Generalized Shallow function	2-100000

Table 1: Test Problems

problem	γ	MSCG	Anderi's
1	0	22/70	21/69*
	$\frac{1}{2}$	24/75	
	1	28/101	
2	0	25/73	21/69*
	$\frac{1}{2}$	31/78	
	1	30/76	
3	0	61/151*	70/201
	$\frac{1}{2}$	failed	
	1	77/163	
4	0	390/1701*	390/1770
	$\frac{1}{2}$	392/1707	
	1	373/1696*	
5	0	21/41	31/51
	1/2	25/69	
	1	49/91	
6	0	71/129*	68/141
	1/2	72/147	
	1	failed	
7	0	4/11	4/12
	1/2	4/10	
	1	4/10*	
8	0	11/201	12/193*
	1/2	12/202	
	1	14/210	
9	0	7/23	8/21
	1/2	6/19*	
	1	7/20	
10	0	8/21	7/22
	1/2	4/11	
	1	7/19*	
11	0	401/991	368/909
	1/2	failed	
	1	338/812	
12	0	17/77	17/78
	1/2	15/70*	
	1	16/78	
totals		951/3273	1017/3536
scores		7	3

Table 2: Function and gradient evaluations

The numerical scores, reported in Table 2, reveal that the new method MSCG displays some improvements over the algorithm in [2] on several problems without the need to store any conditioning matrices. A star appearing next to a score indicates a win on that problem. The last row of Table 2 reports the totals for each method.

The methods tested here implement the same line search strategy with the choices $\rho_1 = 0.0001$ and $\rho_1 = 0.88$ in (18)and (19), respectively. The termination criterion used for the two methods tested here is

$$\|g(x_i)\| \leq 10^{-5}.$$

Both methods were restarted periodically using Powell's test [22] to measure the degree of orthogonality, namely,

$$\left|g_{i+1}^T g_k\right| \geq 0.2\|g_{i+1}\|^2. \tag{20}$$

Whenever (20) is satisfied at step i, the restart is applied with $\beta_i = 0$ in (16), for $\sigma_i = \frac{s_i^T s_i}{s_i^T y_i}$ (see [2]).

5. Conclusions

A new Conjugate Gradient method is developed. The method generates search directions that exploit the multi-step quasi-Newton idea to accelerate convergence of the CG algorithms. The method requires less storage to implement than SCALCG [2]. This save in resources is especially appreciated on large problems.

Other choices for the parameters in (16) are under consideration to determine whether the numerical behaviour of the method can be improved further. There also remains the issue of developing automatic restart criteria that provides appropriate switching among several options similar to what was done in [1]. The global convergence properties of such methods are also under study.

References

[1] M. Al-Baali, New property and global convergence of the Fletcher–Reeves method with inexact line searches, *IMAJ. Numer. Anal.*, **5** (1985), 122-124.

[2] N. Anderi, A scaled BFGS preconditioned conjugate gradient algorithm for unconstrained optimization, *Appl. Math. Letters,* **20** (2007), 645-650.

[3] N. Andrei, New conjugate gradient algorithms based on self-scaling memoryless Broyden–Fletcher–Goldfarb–Shanno method. *Calcolo*, **57**, No 17 (2020); doi:10.1007/s10092-020-00365-7.

[4] C.G. Broyden, The convergence of a class of double-rank minimization algorithms - Part 2: The new algorithm, *J. Inst. Math. Applic.*, **6** (1970), 222-231.

[5] R.H. Byrd, R.B. Schnabel, G.A. Shultz, Parallel quasi-Newton methods for unconstrained optimization, *Math. Programing*, **42** (1988), 273-306.

[6] Y.H. Dai, Y. Yuan, A nonlinear conjugate gradient method with a strong global convergence property, *SIAM J. Optim.*, **10** (1999), 177-182.

[7] J.E. Dennis, R.B. Schnabel, Least change secant updates for quasi-Newton methods, *SIAM Review*, **21** (1979), 443-459.

[8] R. Fletcher, *Practical Methods of Optimization*, 2nd Ed., Wiley, New York (1987).

[9] R. Fletcher, A new approach to variable metric algorithms, *Comput. J.*, **13** (1970), 317-322.

[10] R. Fletcher, C. Reeves, Function minimization by conjugate gradients, *Computer J.*, **7** (1964), 149-154.

[11] J.A. Ford, I.A.R. Moghrabi, Using function-values in multi-step quasi-Newton methods, *J. Comput. Appl. Math.*, **66** (1996), 201-211.

[12] J.A. Ford, I.A.R. Moghrabi, Multi-step quasi-Newton methods for optimization, *J. Comput. Appl. Math.*, **50** (1994), 305-323.

[13] J.A. Ford, I.A.R. Moghrabi, Alternative parameter choices for multi-step quasi-Newton methods, *Optimization Methods and Software*, **2** (1993), 357-370.

[14] J.A. Ford, Y. Narushima, H. Yabe, Multi-step nonlinear conjugate gradient methods for unconstrained minimization, *Comput. Optim. Appl.*, **40** (2008), 191-216.

[15] W. Hager, H.C. Zhang, A new conjugate gradient method with guaranteed descent and an efficient line search, *SIAM J. Optim.*, **16** (2005), 170-192.

[16] M.R. Hestenes, E. Stiefel, Methods of conjugate gradients for solving linear systems, *J. Res. Nat. Bur. Stan. Sec. B*, **48** (1952), 409-436.

[17] Y. Liu, C. Storey, Efficient generalized conjugate gradient algorithms, Part1: Theory, *J. Optim. Theory Appl.*, **69** (1991), 129-137.

[18] I.A.R. Moghrabi, Numerical experience with multiple update quasi-newton methods for unconstrained optimization, *J. of Math. Modeling and Algorithms*, **6** (2007), 231-238.

[19] I.A.R.Moghrabi, Implicit extra-update multi-step quasi-newton methods, *Int. J. Operational Research*, **28** (2017), 69-81.

[20] J.J. Moré, B.S. Garbow, K.E. Hillstrom, Testing unconstrained optimization software, *ACM Trans. Math. Softw.*, **7** (1981), 17-41.

[21] A. Perry, A modified conjugate gradient algorithm, *Oper. Res.*, **26** (1978), 1073-1078.

[22] E. Polak, G. Ribiére, Notesurla convergence de directions conjuguées, *Rev. Francaise Infomat Recherche Operatonelle*, 3e Année, **16** (1969), 35-43.

[23] B.T. Polyak, The conjugate gradient method in extreme problems, *USSR Comp. Math. Phys.*, **9** (1969), 94-112.

[24] M.J.D. Powell, Restart procedures for the conjugate gradient method, *Math. Program.*, **12** (1977), 241-254.

[25] D. Salane, R.P. Tewarson, On symmetric minimum norm updates, *IMA J. Numer. Anal.*, **9**, No 1 (1983), 235-240.

[26] D.F. Shanno, Conditioning of quasi-Newton methods for function minimization, *Math. Comp.*, **24** (1970), 647-656.

[27] D.F. Shanno, K.H. Phua, Matrix conditioning and nonlinear optimization, *Math. Programming*, **14**(1978), 149-160.

[28] D.F. Shanno, On the convergence of a new conjugate gradient algorithm, *SIAM J. Numer. Anal.*, **15** (1978), 1247-1257.

[29] P. Wolfe, Convergence conditions for ascent methods II: Some corrections, *SIAM Rev.*, **13** (1971), 185-188.

[30] L. Wong, , M. Cao, F. Xing, The new spectral conjugate gradient method for large-scale unconstrained optimisation, *J. Inequal. Appl.*, **111** (2020).