Iraqi Journal of Statistical Sciences (13) 2008
p.p. [53-65]

New CG Method for Large-Scale Unconstrained Optimization Based on Nazareth theorem

Khalil . K. Abbo*

Abstract

In this paper we present new conjugate gradient method for computing the minimum value of differentiable real valued function in n variables ,this method derived from Nazareth theorem , which uses the equivalence of CG and Qusi–Newton methods on quadratic function also the descent property and Conjucy conditions are proved and compared with some well know CG method showing considerable improvement .

طريقة جديدة في خوارزميات المتجهات المترافقة لمسائل الامثلية غير المقيدة ذات القياس العالي اعتماداً على نظرية نزرت

الملخص

تم في هذا البحث اقتراح خوارزمية جديدة من المتجهات المترافقة لحساب النهاية الصغرى لدالة الهدف ، وقد تم اشتقاق هذه الطريقة استنادا الى نظرية نزرت وذلك بالاستفادة من تكافؤ طريقة المتجهات المترافقة والطرائق الشبيهة لطريقة نيوتن في الدالة التربيعية كما تم برهان خاصية الانحدار والترافق لهذه الطريقة، وكذلك تمت مقارنة النتائج العددية مع بعض الطرائق المعروفة في هذا المجال .

1-Introduction

A large scale unconstrained optimization problem can be formulated
as the problem of finding a local minimizer of a real valued function
$f : R^n \rightarrow R$ over the space R^n , namely to solve the problem

$$\min \ f(x) \qquad ; \ x \in R^n \qquad ......(1)$$

where the dimension n is large .

The main difficulty in dealing with large scale problems is the fact that effective algorithms for small scale problems do not

*College of Computer Sciences and Mathematics/ Department of Mathematics/University of Mosul
Recieved : 27 /6/ 2007 ____________________Accepted :25/11/2007

necessarily translate into efficient algorithms when applied to solve large problems. Therefore in most cases it is improper to tackle a problem with a large number of variables by using one of the many existing algorithms for small scale case relying on the power growing of the modern computers (see [1] or [2] for a review on the existing methods for small scale unconstrained optimization). A basic feature of an algorithm for large scale problems is a low storage over head needed to make practical its implementation .

Methods for unconstrained optimization differ according to how much information on the function f is available. In the framework of large scale unconstrained optimization it is usually required that the user provides at least subroutines which evaluate the objective function $f(x)$ and its gradient for any $\mathbf{x}$. Throughout, we assume that the function f is twice continuously differentiable i.e the gradient $g(x)=\nabla f(x)$ and Hessian matrix $G(x)=\nabla^2 f(x)$ of the function f exist and are continuous Moreover we denote by $\|.\|_2$ the Euclidean norm .

Most of the large scale unconstrained algorithms (see [3]) are iterative methods which generate a sequence of points according to the scheme

$$x_{k+1}=x_k+\alpha_k d_k \qquad ......(2)$$

where $d_k \in R^n$ is search direction and $\alpha_k \in R$ is a step length obtained by means of a one dimensional search. A basic method for solving (1) can be considered the steepest descent method is obtained by setting in (2)

$$d_k=-g_k \qquad ......(3)$$

This method is based on the linear approximation of the objective function f and hence only first order information is needed. Due to its very limited storage required by a standard implementation, steepest descent method could be considered very attractive in the large scale setting ; moreover the global convergence can also be ensured . However, its convergence rate is only linear and therefore it is too slow to be used .

In 1988 Barzilai and Borwein [4] proposed two point step size gradient (BB) method by regarding

$$H_k = \gamma_k I \qquad ......(4)$$

As an approximation to the Hessian of f at x_k and imposing some quasi – Newton property on H , Denote $v_{k-1} = x_k - x_{k-1}$ and $y_{k-1} = g_k - g_{k-1}$

By minimizing $\|v_{k-1} - H_k y_{k-1}\|_2$ they obtained

$$\gamma_k = \frac{v_{k-1}^T y_{k-1}}{v_{k-1}^T v_{k-1}} \qquad ......(5)$$

With this the BB method is given by the following iteration scheme

$$x_{k+1} = x_k - \frac{1}{\gamma_k} g_k \qquad ......(6)$$

The (BB) method received a great deal of attention for its simplicity and numerical efficiency for well-conditioned problems , the most important features of this method is 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 successfully to find local minimizers of large scale real problems (see [5]). Raydon in [6] proved that for strictly convex function with any variable the (BB) method is globally convergence , despite of these advances of (BB) method on quadratic functions , Fletcher in [7] shows that the method may be very slow on solving some problems .

There are different methods for solving the problem defined in equation (1) corresponding to different ways of choosing d_k in equation (3) , one of the well known effective methods is the Quasi – Newton method in which d_k is defined by

$$d_k = -H_k g_k$$

where H_k is the approximation to the inverse Hessian matrix of the function f at the k-th iteration . There are different ways to update H_k at each iteration (see [8] or [9]), one of the well-known quasi -Newton methods is the DFP method in which is updated by the formula

$$H_{K+1}^{DFP} = H_k + \frac{v_k v_k^T}{v_k^T y_k} - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k} \qquad ......(7)$$

It 's shown that DFP algorithm has quadratic convergence property and H_{k+1} is symmetric , positive definite and hence descent property this leads to global convergence with exact line search with super linear order of convergence, also (see [9]) H_{k+1} satisfies quasi – Newton condition

$$H_{k+1} y_k = v_k \qquad ......(8)$$

The main disadvantage of the quasi-Newton methods is storing matrix .

2- Non-linear Conjugate Gradient methods (CG).

CG uses the analytic derivative of f , defined by g_k. A step along the current negative gradient vector is taken in the first iteration ; successive directions are constructed so that they form a set of mutually conjugate vectors with respect to the Hessian. At each step, the new iterate is calculated from eq (2) and the search directions are expressed recursively as

$$d_{k+1} = -g_{k+1} + \beta_k d_k \qquad ......(9)$$

where β_k is scalar and step length α_k is required to satisfy the strong Wolfe conditions

$$f(x_k + \alpha_k d_k) - f(x_k) \le \delta \alpha_k g_k^T d_k \qquad ......(10)$$

$$\left| g(x_k + \alpha_k d_k)^T d_k \right| \le -\sigma g_k^T d_k \qquad ....(11)$$

where $0 < \delta < \sigma < 1$

For a general function, however different formula for scalar β_k result in distinct non-linear conjugate gradient methods and for quadratic function all β_k are equivalent. Several famous formulas β_k are the Fletcher- Reeves (β_{FR}), Polak Ribiere (β_{PR}) Hestenes- Stiefel (β_{HS})

And Yu- Hong (β_{yH}) and Perry (β_{pr})(see [9] and [10]) which are given

$$\beta_k^{FR} = \frac{g_k^T g_k}{g_{k-1}^T g_{k-1}} \qquad \text{......(12)}$$

$$\beta_k^{PR} = \frac{g_k^T y_{k-1}}{g_{k-1}^T g_{k-1}} \qquad \text{......(13)}$$

$$\beta_k^{HS} = \frac{g_k^T y_{k-1}}{d_{k-1}^T y_{k-1}} \qquad \text{......(14)}$$

$$\beta_k^{YH} = \frac{g_k^T g_k}{d_{k-1}^T y_{k-1}} \qquad \text{......(15)}$$

$$\beta^{pr} = \frac{(y_{k-1} - v_{k-1})^T g_k}{y_{k-1}{}^T v_{k-1}} \qquad \text{......(16)}$$

In practical computation , the HS method resembles the PR method (see [11] or [12]), both methods are generally believed to be two of the most efficient conjugate gradient methods .

Most of the recent work in nonlinear CG methods has focused on global convergence properties and on the design of new line search strategies . The analysis for the FRCG method is simpler ,it shown in [13] that if the line search satisfies the strong Wolfe conditions then the Fletcher-Reeves method is globally convergent . The same result is proved in [14] for all CG methods with line search satisfying the strong Wolfe conditions and with any β_k such that $0 \le \beta_k \le \beta_k^{FR}$.

The analysis is taken one step further in [15] ,where it is shown that global convergence is obtained for any method with

$$|\beta_k| \le \beta_k^{FR} \qquad \text{......(17)}$$

A major drawback of non – linear CG methods is that the search direction tends to be poorly scaled , and line search typically several function evaluations to obtain an acceptable step length α_k. This is in sharp constant with quasi-Newton method which accepts the unit step length most of the time (see [16]). Non-linear CG methods would therefore be greatly improved if we could find a means of properly scaling d_k . Many studies have suggested search directions of the form

$$d_k = -H_k g_k + \beta_k d_{k-1} \qquad ......(18)$$

where H_k is simple symmetric and positive definite matrix of satisfying eq (8) .However ,if H_k requires several vectors of storage , the economy of the non linear CG iteration disappears. So far all attempts to derive an efficient method of the form (18) have been unsuccessful .

Nazareth in [17] has pointed out a close relationship that exists between the CG algorithm and Quasi-Newton algorithms , in fact he shows that for quadratic function with exact line search CG and DFP methods generates the same sequence $\{x_k\}_{k=1}^{\infty}$ and same directions d_k for all k i.e

$$d_k^{CG} = d_k^{DFP} \qquad ......(19)$$

We can use this equivalence of CG and DFP methods to deduce new CG method.

In this paper we attempt to combine QN and CG methods (in different way from (18)) to deduce new CG methods which use Quasi-Newton method implicitly .

3- New proposed CG Algorithms

From Nazareth theorem we have $d^{DFP} = d^{CG}$ then

$$-H_k^{DFP} g_k = -\theta g_k + \beta_k d_{k-1} \qquad ......(20)$$

Where θ is scalar $(0 < \theta_k \le 1)$. Multiply eq (20) by y_{k-1} we get

$$-H_k y_{k-1}^T g_k = -\theta_k y_{k-1}^T g_k + \beta_k y_{k-1}^T d_{k-1} \qquad ......(21)$$

use quasi-Newton condition defined in equation (8) then we can write (21) as

$$-v_{k-1}^T g_k = -\theta_k y_{k-1}^T g_k + \beta_k y_{k-1}^T d_{k-1}$$

$$\boldsymbol{\beta_k^{New} = \frac{(\theta_k y_{k-1} - v_{k-1})^T g_k}{y_{k-1}^T d_{k-1}}} \qquad \textbf{......(22)}$$

Where $\theta_k = 1$ or

$$\theta_k = \frac{v_k^T v_k}{v_{k-1}^T v_{k-1} + y_{k-1}^T v_{k-1}} \qquad \text{.....(23)}$$

where θ_k $(\boldsymbol{\theta_k > 0 \ \forall\, k})$ in (22) due to Abbo (see [18]), therefore the search direction for new 1 can be written as

$$\mathbf{d_{k+1} = -\theta_{k+1} g_{k+1} + \beta_k^{New} d_k} \qquad \textbf{......(9)}$$

$$d_k = -\theta_k g_k + \frac{(\theta_k y_{k-1} - v_{k-1})^T g_k}{y_{k-1}^T d_{k-1}} d_{k-1} \qquad \text{......(24)}$$

It's clear that if exact line search used $\theta_k = 1$, then β_k in (22) reduced to β_k^{HS}, on the other hand if d_{k-1} is replaced by the $\frac{1}{\alpha_{k-1}} v_{k-1}$ in the denominator and again with $\theta_k = 1$we get Perry CG method (see [19]).

Out Line of the Algorithem(New)

step(1): $k = 0$; choose $x_o \in R^n; \varepsilon > 0;\ d_o = -g_o$

step(2): if $\|g_k\| < \varepsilon$ stop, else goto step(3)

step(3): Compute $\alpha_k$by (inexact line search procedure) with Wolfe conditions

step(4): $x_{k+1} = x_k + \alpha_k d_k$

step(5): Compute g_{k+1}, y_k, v_k

step(6): Compute search direction from eq(24) with $\theta = 1$ or θ as definde in (23)

step(7): $k = k + 1$ goto step 2

To prove the descent property of the New algorithm we have two cases

(1): if exact line search is used then we see from (24) that for each $k \geq 1$, the directional derivative of f at x_k along direction d_k is given by

$$g_k^T d_k = -\theta_k g_k^T g_k + \frac{(\theta y_{k-1} - v_{k-1})^T g_k}{y_{k-1}^T d_{k-1}} g_k^T d_{k-1}$$

then we have for any $k \geq 1$

$$g_k^T d_k = -\theta \|g_k\|^2 < 0$$

(2) if inexact line search is used then

$$g_k^T d_k = -\theta_k g_k^T g_k + \frac{(\theta_k y_{k-1} - v_{k-1})^T g_k}{d_{k-1}^T y_{k-1}} g_k^T d_{k-1}$$

$$= -\theta_k \|g_k\|^2 + \frac{(\theta_k y_{k-1} - v_{k-1})^T d_{k-1}}{d_{k-1}^T y_{k-1}} g_k^T g_k$$

$$= -\theta_k \|g_k\|^2 + \theta_k \|g_k\|^2 - \frac{\frac{1}{\alpha} v_{k-1}^T v_{k-1}}{\frac{1}{\alpha} v_{k-1}^T y_{k-1}} \|g_k\|^2$$

$$= -\frac{v_{k-1}^T v_{k-1}}{v_{k-1}^T y_{k-1}} \|g_k\|^2 < 0$$

It is well known that $v_{k-1}^T y_{k-1} > 0$ for CG and QN methods see[9] therefore the search direction defined in (24) is descent (one can use $v_{k-1}^T y_{k-1} > 0$ as restart to forcing descent property to avoid effect round of error, if inexact line search employed .

The conjucy condition is hereditary from HSCG if exact line search is used and from Perry CG method if inexact line search is used, therefore the global convergence is a consequence of descent, conjucy and Wolfe conditions if further we assume that the level set

$L = \{x : f(x) \leq f(x_o)\}$ is bounded.

Numerical Results

We present the numerical results for HSCG, Perry CG and New1(with $\theta=1$ and θ as defined in (23)) methods for some well known test functions taken from [20], these algorithms are coded in double precision FORTRAN language. The criteria for stopping the iteration is

$$\|g_k\| < 10^{-6}$$

The line search procedure used in this work is the Birgin and Mortaniz [21] method with initial step size equal one in all methods. Also Wolfe conditions are used for accepting step size the complete set of results are given in table (a) with $1000 \le n \le 5000$ and table (b) with $6000 \le n \le 10000$. In tables (a)and (b) we present the comparison results of HSCG, Perry CG and New1methods for different dimensions consisting number of iteration NOI , number of functions evolutions NOF are compared it's shown that considerable improvement over the other methods

Table(a) comparison CG methods for 1000<n<5000

Test Functions	N	HS		Pery		New 1 $\theta = 1$		New 1 $\theta = \frac{v^T v}{v^T v + y^T v}$	
		NoI	Nof	NoI	Nof	NoI	Nof	NoI	Nof
Extended Trigonometric	1000	52	101	24	52	19	41	16	34
Extended Rosenbrock	=	21	63	31	91	27	82	30	78
Extended Beal	=	19	66	16	35	17	36	13	32
Perturbed Quadratic	=	200	375	230	451	187	375	221	431
Diagonal 2	2000	244	479	214	439	226	450	260	513
Generalized Tridiagonal	=	30	101	47	118	25	91	28	93
Extended Tridiagonal 1	=	14	30	16	29	13	27	15	32
Extended 3 Exponential Terms	=	16	46	14	47	8	18	16	30
Generalized Tridiagonal 2	3000	41	158	40	73	115	348	45	81
Generalized Rosenbrock	=	27	58	18	34	21	75	15	27
Generalized PSCI	=	84	261	90	244	81	199	98	208
Extended PSCI	=	26	53	22	40	39	69	28	49
Extended Powell	4000	24	78	29	97	15	54	23	79
Full Hessian FH2	=	259	619	337	640	259	519	242	644
Extended Block Diagonal BDI	=	164	559	56	133	90	267	41	107
Extended Maratos	=	89	507	71	462	67	322	64	274
Extended Cliff	5000	426	853	430	849	426	853	430	749
Quadratic Diagonal Perturbed	5000	14	81	14	76	11	33	16	69
Extended Wood	5000	62	192	48	173	51	209	48	169
Extended Quadratic Penalty	5000	56	112	66	229	47	162	48	330
		1868	4792	1815	4312	1744	4230	1697	4029

Table (a1) Percentage of improving the New 1 within $\mathbf{1000 \le n \le 5000}$

Tools	HSCG	Perry CG	New $\theta = 1$	New $\theta = \frac{v^T v}{v^T v + y^T v}$
NOI	100%	97%	93%	90%
NOF	100%	89%	88%	84%

Table(b) comparison CG methods for 6000<n<10000

Test Functions	N	HS		Pery		New 1 $\theta=1$		New 1 $\theta=\frac{v^Tv}{v^Tv+y^Tv}$	
		NoI	Nof	NoI	Nof	NoI	Nof	NoI	Nof
Extended Trigonometric	6000	71	139	53	118	46	109	42	103
Extended Rosenbrock	=	27	72	29	96	28	87	27	81
Extended Beal	=	16	33	14	33	17	35	13	32
Perturbed Quadratic	=	230	467	241	492	227	452	223	432
Diagonal 2	7000	297	576	248	541	237	512	221	520
Generalized Tridiagonal	=	24	61	35	82	22	58	21	52
Extended Tridiagonal 1	=	18	37	15	27	15	26	16	34
Extended 3 Exponential Terms	=	21	58	14	47	11	28	9	21
Generalized Tridiagonal 2	8000	52	149	45	128	97	192	41	123
Generalized Rosenbrock	=	41	101	37	92	26	70	18	41
Generalized PSCI	=	90	222	85	190	65	162	68	167
Extended PSCI	=	39	70	31	65	39	79	30	57
Extended Powell	9000	24	78	19	71	16	64	21	66
Full Hessian FH2	=	398	797	380	769	291	686	252	671
Extended Block Diagonal BDI	=	202	421	181	397	164	368	150	342
Extended Maratos	=	91	509	77	462	72	453	67	441
Extended Cliff	10000	482	897	501	920	490	911	482	900
Quadratic Diagonal Perturbed	10000	28	110	20	59	18	48	21	52
Extended Wood	10000	63	195	56	182	51	209	62	195
Extended Quadratic Penalty	10000	87	213	68	146	63	131	58	132
		2301	5205	2149	4917	1995	4680	1832	4462

Table (a1) Percentage of improving the New 1 within **$6000 \le n \le 10000$**

Tools	HSCG	Perry CG	New $\theta=1$	New $\theta=\frac{v^Tv}{v^Tv+y^Tv}$
NOI	100%	94%	87%	80%
NOF	100%	94.5%	90%	85%

REFERENCES

[1] Fletcher, R. (1994). " An overview of unconstrained optimization , in Algorithms for continuous optimization". The state of the art , E. spedicato , ed., Kluwer Academic publishers .

[2] Nocedal, J**.** (1992) "Theory and algorithms for unconstrained optimization" Acta Numerica , 1992.

[3] Rama , M. (1999) " Large scale unconstrained optimization " Technical Report 10-99. Encylopedia of optimization C. Floudas and P. Pardalos editors Kluwer Academic publishers.

[4] Barzilai , J. and Borwein **, (1998).** M. " Two points step size Gradient Method " IMA J. Number Anal.,8.

[5] Luengo, F. and Raydan, M. (2003). " Gradient Method with Dynamical Retards for large scale optimization problems" Electronic Transactions on Numerical Analysis Vol (16).

[6] Raydan, M. (1993). "On the Barizilai – Browein choice of the step length for the gradient method " IMA J. Number Anal.,13.

[7] Fletcher ,R. (2001). "On the Barizilai - Browein method " Numerical Analysis Report NA 1207,October.

[8] Edwin , K. and Stanislaw , H. (2001). "An Introduction to optimization" John Wiley& sans , Inc.

[9] Fletcher, R.(1987). "Practical Methods of optimization" (**2nd** Edition). John Wiley Chichester.

[10] Dai Y. H. and Yuan,Y. (1999). "A Nonlinear Conjugate Gradient Method with A strong Global Convergence Property " STAM Journal on optimization ,10 (1).

[11] Hiroshi, Y. and Masahiro , T. (2004). " Global Convergence Properties of Nonlinear Conjugate Gradient Methods with Modified secant condition" ,Computational optimization and Applications 28.

[12] Armand, P. (2006) "Modification of the Wolf line search Rules to satisfy the descent Condition in Polak – Ribiere – Polyak CG method " Journal of optimization theory and Applications Springer science + Business Media. Inc.

[13] AL – Baali, M. (1985) "Descent Property and global Convergence of the Fletcher - Reeves method with inexact line search " IMA Journal of Numerical Analysis 5.

[14] Touati, A. and Story, C. (1990). "Efficient hybrid Conjugate Gradient techniques" Journal of Optimization Theory and Applications 64.

[15] Gillbert, J.C. and Nocedal, J. (1990) " Global Convergence Properties of conjugate gradient methods for optimization ", SIAM Journal on optimization , 2,(1).

[16] Nocedal , J. (1996). " Large scale unconstrained optimization "Tech. Report G.1.6, Numerical Analysis , Department of Electrical Engineering and computer Science North western University .

[17] Nazareth , L. (1997). "A relation ship between the Qusi – Newton and CG algorithms", AMD Tech. Memo 282, Argonne National Laboratory .

[18] Abbo , K. (2007). "Modifying of Barzilai and Borwein Method for solving Large scale unconstrained optimization problems "IRQ.J.S.S, Vol. (7), No.11.

[19] Perry, A. (1978). "A Modified Conjugate Gradient Algorithm " Operations Research 26,

[20] Bongartz, I.; Conn, A.; Gould N. and Toint, P. (1995). "CUTE: Constrained andUnconstrained testing environments", ACM Trans. Math. Software.

[21] Birgin, G. and Martinez M. (2001). "A spectral Conjugate gradient method for unconstrained optimization", Applied Mathematics and Optimization 42.