Investigation on Scaled CG-Type Algorithms for Unconstrained Optimization

Abbas Y.AL-Bayati Khalil K. Abo Salah G. Shareef
profabbasalbayati@yahoo.com
College of Computer sciences and Mathematics
University of Mosul, Iraq

Received on: 05/03/2006 Accepted on: 16/08/2006

ABSTRACT

In this paper, we describe two new algorithms which are modifications of the Hestens-stiefl CG-method. The first is the scaled CG-method (obtained from function and gradient-values) which improves the search direction by multiplying to a scalar obtained from function value and its gradient at two successive points along the iterations. The second is the Preconditioned CG-method which uses an approximation at Hessein of the minimizing function. These algorithms are not sensitive to the line searches. Numerical experiments indicate that these new algorithms are effective and superior especially for increasing dimensionalities.

Keywords: Unconstrained Optimization, Conjugate gradient algorithm, Hestens-stiefl method, Hessein matrix.

التقصي في خوارزميات التدرج المترافق الطيفي في الامثلية غير المقيدة

صلاح غازي شريف خليل خضر عبو عباس يونس البياتي
كلية علوم الحاسوب والرياضيات/جامعة الموصل/العراق

تاريخ استلام البحث: 2006/03/05 تاريخ قبول البحث: 2006/08/16

الملخص

تم في هذا البحث اقتراح خوارزميتين جديدتين لتحسين خوارزمية المتجهات المترافقة لطريقة (HS), الأولى تعتمد على ضرب متجهات البحث في (HSCG) بقياسي (γ_k)يمكن (حصوله من قيمة الدالة وقيمة المشتقة الأولى) عند نقطتين متتابعتين اما الخوارزمية الثانية فتم ضرب متجه البحث بمصفوفة تعتبر تقريبا جديدا لمصفوفة هيسي لدالة الهدف. وهاتان الخوارزميتان غير حساستين للبحث الخطي. وتشير التجارب العددية لبعض دوال الاختبار إلى كفاية هذه الخوارزميات مقارنة بمثيلاتها عند الأبعاد الكبيرة.

الكلمات المفتاحية: الأمثلية غير المقيدة، خوارزمية التدرج المترافق، طريقة HS، مصفوفة هيسي.

1. Introduction:

Unconstrained Optimization Problems expressed as:

$$\min f(x); x \in R^n \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Where $f:R^n \longrightarrow R$ is twice differentiable real valued function, is one of the most active areas in optimization community. virtually appearing in every human activity.

For solving these problems many efficient methods have been suggested. Excellent Presentations of these methods can be found, for example (Fletcher, 1987; Gill et-al 1981; Boyd and Vavdenberhe, 2003; Nocedal, 1992; Edwin 2001). The most useful algorithms classify in: The Conjugate gradient and its variants, Newton method and its extensions; the DFP variable metric method, many different QN methods. All these methods are iterative and consider iterations of the form:

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

where d_k is a descent direction i.e.

$$d_k^T g_k \prec 0 \qquad \text{……...……(3)}$$

Where $g_k = \nabla f(x_k)$ and α_k is a step length obtained by line search.Conjugate Gradient methods consider the search directions as:

$$d_{k+1} = -g_{k+1} + \beta_k d_k,\ k \geq 1 \qquad \text{…….……(4)}$$

where the scalar β_k is chosen in such a manner that the method reduces to the linear Conjugate Gradient when the function is quadratic and line search is exact. The rest of the methods define the search directions by:

$$d_k = -G_k^{-1} g_k, \qquad \text{…….……..(5)}$$

where G_k is a non-singular symmetric matrix. Mainly the matrix G_k is selected as $G_k = I$ (identity matrix which gives the steepest descent method), $G_k = \nabla^2 f(x_k)$ (the Newton method) or an approximation of the Hessian $\nabla^2 f(x_k)$.

Different modifications are made to the CG-algorithm in different ways (see for example Hu and Story, 1990; Fletcher, 1993; Al-Baali 1985), most of these modifications are made to the search directions to improve then.

We end this general introduction by content of this paper which is organized as follows: In section(2) we review the Conjugate gradient, QN-methods and their Combinations, section(3) contains the development of the new algorithms, the last section includes the numerical results.

2. Review of the methods:

2-1 Conjugate Gradient Methods:

Non-linear Conjugate gradient (CG) is one of the most useful and the earlist techniques for solving Large-scale non-linear optimization problems.

Many variants of this original scheme have been proposed, and some are widely used in practice, CG-methods only use the first order derivatives information of the objective function and need not update the Hessian matrix at each iteration. First, these are used to solve the general unconstrained optimization problems by Fletcher and Reeves (1964).

Conjugate gradient methods depend on the fact that for quadratic, if we search along a set of n mutually conjugate directions $d_k, k=1,2,...,n$ (by Conjugate directions we mean that $d_i^T G d_j = 0$, $\forall i \neq j$, where G is $n \times n$ positive matrix), then we will find the minimum in at most n-steps if line searches are exacts i.e.

$$d_k^T g_{k+1} = 0 \qquad \text{……..(6)}$$

Moreover, if we generate this set of directions by known gradients then each direction can be simply expressed as:

$$\left.\begin{aligned} d_1 &= -g_1 \\ d_{k+1} &= -g_{k+1} + \beta_k d_k \ , \ k \geq 1 \end{aligned}\right\} \qquad \text{....…….....…….(7)}$$

Where

$$\beta_k = \frac{g_{k+1}^T y_k}{d_k^T y_k} \qquad \text{…….…….....…….(8)}$$

And

$$y_k = g_{k+1} - g_k \qquad \text{…..…….……(9)}$$

β_k given in equation (8) is called the Hestenes-stifel formula; clearly there are other different forms of β_k such as Polak-Ribeier formula which is derived from β_k in equation (8) as follows:

$$\frac{g_{k+1}^T y_k}{d_k^T y_k} = \frac{g_{k+1}^T (g_{k+1} - g_k)}{d_k^T (g_{k+1} - g_k)} = \frac{g_{k+1}^T g_{k+1} - g_{k+1}^T g_k}{d_k^T g_{k+1} - d_k^T g_k}$$

Assuming equation (6) and considering :

$$d_k = -g_k + \beta_{k-1} d_{k-1}$$

(see Edwin 2001) we get $d_k^T g_k = -g_k^T g_k$

Hence $\frac{g_{k+1}^T y_k}{d_k^T y_k} = \frac{g_{k+1}^T y_k}{g_k^T g_k} = \beta_{PR}$ ……...…(10)

The Fletcher Reeves formula can be derived from β_{PR} by considering $g_{k+1}^{T} g_k = 0$ to obtain $\beta_{FR} = \frac{g_{k+1}^T g_{k+1}}{g_k^T g_k}$ ……..(11), and there are many other forms. All these β_k are equivalent on quadratics with exact line searches and starting with steepest descent direction, but when extended to general non-linear functions, the conjugate gradient algorithms with different β_k are quite different in effacing and are considered to be not so efficient as the QN algorithms in general. The focus in this paper to the β_k given in equation (8) since its original CG-method.

Finally, in many implementations of Conjugate gradient methods, The iteration (7) is restarted every n or (n+1) steps setting β_k equal to zero i.e. taking steepest descent step. This ensures global convergence (Nocedal, 1992). However, different restarts are introduced (see Fletcher, 1987). One of the well-known restarts given by Powell(Powell 1977) is:

$\left| g_{k+1}^T g_k \right| \geq 0.2 \left\| g_k \right\|^2$ ……(12)

This criterion will be used later in our suggested algorithms.

2-2 Quasi-Newton Methods:

Quasi-Newton methods are probably the most popular general purpose algorithms for unconstrained optimization problems. Many QN-methods are advantageous due to their fast convergence and absence of second order derivatives computation.

For the QN-methods assume that at the *kth* iteration at approximation point x_k and $n \times n$ matrix H_k are available. Then the methods proceed by generating a sequence of approximation points via the equation:

$$x_{k+1} - x_k + \alpha_k d_k$$

and

$$d_{k+1} = -H_k g_k \qquad \text{…….....……(13)}$$

Where H_k is an approximation of G_k^{-1}which is corrected or updated form iteration to iteration, In general, H_k is symmetric and positive definite, there are different choices of H_k (see Fletcher, 1987), we list here some most popular forms (Frandsen et. al. 2004)

$$H_{k+1}^{SR} = H_k + \frac{(v_k - H_k y_k)(v_k - H_k y_k)^T}{(v_k - H_k y_k)^T y_k} \quad \text{…….……(14)}$$

is called Rank one Correction formula,

where $v_k = x_{k+1} - x_k$ and y_k as in equation (9)

$$H_{k+1}^{DFP} = H_k + \frac{v_k v_k^T}{v_k^T y_k} - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k} \quad \text{…….……(15)}$$

is the DFP formula

$$H_{k+1}^{BFGS} = H_k + \left[1 + \frac{y_k^T H_k y_k}{v_k^T y_k}\right]\frac{v_k v_k^T}{v_k^T y_k} - \left[\frac{v_k y_k^T H_k + H_k y_k v_k^T}{v_k^T y_k}\right] \text{……..(16)}$$

All three forms satisfy the Quasi-Newton Condition

$$y_k = G_k v_k \text{ or } H_{k+1} y_k = v_k$$

and maintains positive definite matrices if H_0 is positive.

We conclude our discussion of the QN methods with the following theorems:

Theorem (1): If the DFP algorithm is applied to the quadratic with Hessian $G = G^T$ we have

$$H_{k+1} y_i = v_i \quad \text{…….....…… (17)}$$

Theorem (2): Suppose that $g_k \neq 0$. In the DFP algorithm, if H_k is positive definite then so H_{k+1}. For proof of the theorems (1) and (2) (see Dixon, 1972)

2-3 Preconditioned CG algorithm (PCG):

The Preconditioned CG methods (PCG) first appeared in paper by Axelsson (Axelsson, 1972). It was developed with object of accelerating the convergence of the CG-method by a transformation of variables while keeping the basic properties of the method. Such transformation was introduced by Allwright (Allwright, 1972), the symmetric positive definite matrix H can be factored in various ways for example as $H = LL^T$ where L is lower triangular and non-singular (for more detail see Allwright, 1972).

Buckley (Buckley, 1978. a and b) introduced an algorithm in which conjugate gradient and quasi-Newton search directions occur together and which can be interpreted as a conjugate gradient algorithm with changing

metric. Many authors have extended this type of algorithms (see for example Al-Bayati, 1996).

The search direction to the preconditioned (PCG) method is defined by:

$$\left.\begin{aligned} d_1 &= -H_1 g_1 \\ d_{k+1} &= -Hg_k + \beta_k d_k \text{ for } k \geq 1 \end{aligned}\right\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(18)$$

$$\beta_{HS} = \frac{g_{k+1}^T Hy_k}{d_k^T Hy_k}$$

Where H is one of the forms given in equations (7) or (8) or (9). In this paper, our focus is to the H^{PDF}.

3. Development of two suggested Algorithms

3-1 Scaled CG-method New1 (say)

One of the reasons for inefficiency of conjugate gradient algoriths is that non of the β_k takes into consideration the effect of inexact line searches (Hu and Story, 1990). In order to do this and find an optimal β, Liu and Story (1991) introduced an algorithm that finds an optimal (β_{LS}) by solving a quadratic function f as:

$$F(x_{k+1}, d_{k+1}) = \min_{\alpha} f(x_{k+1} + \alpha d_{k+1}) \qquad \ldots\ldots\ldots\ldots\ldots\ldots(19)$$

and

$$\min_{\beta} F(x_{k+1}, -g_{k+1} + \beta_k d_k) \qquad \ldots\ldots\ldots\ldots\ldots\ldots(20)$$

an then takes $d_{k+1} = \alpha^*(-g_{k+1} + \beta_{LS} d_k)$

Where α^* and β_{LS} solve (19) and (20), respectively, the major drawback in Liu and Story CG algorithm is solving equation (20) at each iteration.

In this paper, a new form of CG-method presented with line search and so it develops conjugate search directions. This new approach will find the minimum of a $n-$dimensional quadratic function in at most ($n+1$) function and gradient evaluations.

The motivation of the iteration

$$x_{k+1} = x_k + \alpha_k d_k \ , \ \alpha_k \succ 0 , \qquad \ldots\ldots\ldots\ldots\ldots\ldots(21)$$

is that the search direction d_i is chosen to ensure that the point x_{i+1} is a minimize of the model

$f(x_k+\alpha_k d_k)=f(x_k)+g_k^T v_k+\frac{1}{2}v_k^T G_k v_k$, ……..…..…...……..(22)

where $v_k=x_{k+1}-x_k=\alpha_k d_k$, ……...……(23)

or $f(x_{k+1})=f(x_k)-\alpha_k g_k^T g_k+\frac{1}{2}\alpha_k^2 g_k^T \nabla^2 f(z) g_k^{/}$, ……...……..(24)

where $z\in[x_k, x_{k+1}]$. If the distance between x_k and x_{k+1} is small enough we can choose $z=x_{k+1}$ and consider $\gamma(x_{k+1})I$ as an approximation of the $\nabla^2 f(x_{k+1})$, where $\gamma(x_{k+1})\in R$. This is an approximation of the Hessian at a point x_{k+1} is computed using the local information from point x_k therefore we can write

$$\gamma(x_{k+1})=\frac{2}{g_k^T g_k}\frac{1}{\alpha_k^2}[f_{k+1}-f_k+\alpha_k g_k^T g_k] \quad \text{……..……(25)}$$

where the step size α_k is satisfying line search conditions (Wolfe-Powell conditions) such as

$f(x_k+\alpha_k d_k)\le f(x_k)+\rho\alpha_k g_k^T d_k$ and $\left|\nabla f(x_k+\alpha_k d_k)^T d_k\right|\le-\sigma\left|\nabla f(x_k)^T d_k\right|$ ……(26)

where $0\prec\rho\prec\sigma\prec 1$

Then we take $\gamma(x_{k+1})I$ as an approximation of G. It is clear that if $\gamma(x_{k+1})\succ 0$ this approximation will be positive definite hence to complete the method.

We must consider the situation when $\gamma(x_{k+1})\prec 0$, i.e. if

$$f(x_{k+1})-f(x_k)+\alpha_k g_k^T g_k\prec 0$$

In this case we can change the step size α_k as $\alpha_k+\eta_k$ s.t.

$f(x_{k+1})-f(x_k)+(\alpha_i+\eta_k)g_k^T g_k\succ 0$ ……(27)

to get a value for η_k. Let us select a $\delta_k\succ 0$ small enough and consider

$$\eta_k=\frac{1}{g_k^T g_k}[f(x_k)-f(x_{k+1})-\alpha_k g_k^T g_k+\delta] \quad \text{……(28)}$$

then the new value of $\gamma(x_{k+1})$ can be computed from

$$\gamma(x_{k+1})=\frac{2}{g_k^T g_k}\frac{1}{(\alpha_k+\eta_k)^2}[f(x_{k+1})-f(x_k)+(\alpha_k+\eta_k)g_k^T g_k] \quad \text{……....(29)}$$

then the new algorithm can be obtained by multiplying the search direction to scalar $\frac{1}{\gamma(x_{k+1})}$ as follows:

The outlines of the Self-Scaling CG-method (new1):

Step (1): Set initial point x_1 and scalar ε .

Step (2): For $k=1$ set $d_1=-g_1$

Step (3): For $k\geq 1$ compute $x_{k+1}=x_k+\alpha_k d_k$ where α_k is obtained by line search procedure

Step (4): If $\|g_k\|\leq\varepsilon$ stop. Otherwise go to step(5).

Step (5): Compute $v_k=x_{k+1}-x_k, y_k=g_{k+1}-g_k$.

Step (6): Compute $\gamma(x_{k+1})$ from equations (25) or (28 and 29).

Step (7): Set the new β_k to

$$\beta_k=\frac{1}{\gamma_{k+1}}\frac{g_{k+1}^T y_k}{d_k^T y_k}$$

Step (8): set $d_{k+1}=-\frac{1}{\gamma_{k+1}}g_{k+1}+\beta_k d_k$

Step (9): If restart satisfied (restart Powell 1977) go to step (2) else k=k+1 go to step (3)

3-2 Self-Scaling PCG method (New2):

In this section, a new PCG method for solving unconstrained optimization problems is proposed .

This new PCG algorithm considered here has an additional property of being invariant under scaling of the function or of its variables where the objective function is twice continuously differentiable and search direction is descent i.e. $g_k^T d_k\prec 0$ also we assume that line search is exact i.e. $g_k^T d_k=0$.

Let $d_k=-H_k g_k+\beta_k d_{k-1}$ where $\beta_k=\beta(y_k d_k, H_k g_{k+1})$. where the matrix H_k is an approximation of G^{-1} the inverse of Hessian of the objective function $f(x)$.

One important feature of PCG method is the choice of H_k . The method requires H_k to be positive definite to deduce directions.

Let $H_k=\gamma_{k+1}I$ as an approximation to the inverse Hessian where I and identity matrix and γ_{k+1} computed from (25), if $\gamma_{k+1}\succ 0$ and from (29) if $\gamma_{k+1}\prec 0$ and satisfy the condition

$$H_k=\arg\min_{\gamma}\|H_k y_k-v_k\|_2 \qquad \text{.....(30)}$$

Where $\|.\|_2$ usual Euclidian norm, then the new2 algorithm can be organized as follows:

The outlines of the Self-Scaling PCG method (new2):

Step(1): Set $x_1, \varepsilon, H_1 = I$.

Step(2): For $k = 1$ set $d_1 = -H_1 g_1$.

Step(3): Set $x_{k+1} = x_k + \alpha_k d_k$ for $k \geq 1$, where α_k optimal step size.

Step(4): If $\|g_k\| \leq \varepsilon$ stop. Otherwise

Check if restart equation (12) is satisfied then set $x_k = x_1$ go to step (2).

otherwise go to step (5).

Step(5): Compute $y_k = g_{k+1} - g_k; v_k = x_{k+1} - x_k$ and γ_k form equation (25).

Step(6): Compute $H_k = \gamma_k I$ satisfying (30) and $\beta_k = \dfrac{g_{k+1}^T H_k y_k}{d_k^T y_k}$.

Step(7): Set $d_{k+1} = -H_k g_{k+1} + \beta_k d_k$ go to step (3).

4. Numerical Results:

All the algorithms described in this paper namely:

1. The standard HSCG method.
2. Preconditioned CG with (DFP and BFGS).
3. New1 and New2 proposed algorithms are coded in double precision FORTRAN 90. The complete set of results is given in Table (1) and Table (2). In our numerical comparison, the number of function evaluations NOF and number of iterations NOI are considered. The actual convergence criterion employed was $\|g_{k+1}\| \prec 1 \times 10^{-5}$ for all the algorithms. Well-known test functions with different dimensions n are employed in this comparisons.

Table (1a) Comparison of HSCG and New1

Test Fun.	N	HSCG		New1	
		NOI	NOF	NOI	NOF
Dixon	4	13	28	13	28
Powell (4)	4	50	114	46	96
Rosen	4	28	68	23	57
Cubic	4	16	42	15	(36)
Wood	4	31	67	21	45
Dixon	10	22	46	21	44
Powell (4)	20	34	78	25	53

Rosen	20	23	58	23	59
Cubic	20	14	37	11	32
Wood	20	52	107	31	66
Powell (4)	100	129	263	67	(148)
Rosen	100	23	58	22	57
Cubic	100	14	37	11	32
Wood	100	69	140	85	174
Powell (4)	500	458	921	68	(165)
Rosen	500	23	58	22	57
Cubic	500	14	37	12	35
Wood	500	69	140	86	176
Powell (4)	1000	558	1121	70	(160)
Rosen	1000	23	58	22	57
Cubic	1000	14	37	12	35
Wood	1000	70	142	90	184
Total		1747	3657	786	1786

(1b) percentages of improving the New1 method

Tools	**HSCG method**	**New1 method**
NOI	100%	55%
NOF	100%	59%

Clearly there is an improvement of 45% in NOI and 41% in NOF for our new proposed algorithms.

Table (2a) Comparison of PCG methods with DFP , BFGS and New1

Test Fun.	**N**	**PCG with DFP**		**PCG with BFGS**		**New2**	
		NOI	**NOF**	**NOI**	**NOF**	**NOI**	**NOF**
Powell (4)	4	22	79	21	86	42	104
Cubic	4	19	88	18	51	11	32
Dixon	4	10	31	14	30	13	28
Wood	4	54	149	37	109	21	45
Rosen	4	23	63	34	87	30	81
Powell (4)	20	36	135	38	123	30	83
Cubic	20	37	93	18	51	11	32
Dixon	10	31	90	22	47	21	44
Wood	20	130	353	84	243	44	94
Rosen	20	68	187	34	87	30	81
Powell (4)	100	82	387	71	197	31	94
Cubic	100	48	153	18	51	11	32
Wood	100	243	861	251	774	85	147

Rosen	100	155	389	34	87	30	81
Powell (4)	500	54	284	50	148	36	113
Cubic	500	76	389	18	51	11	32
Wood	500	288	891	200	801	86	176
Rosen	500	178	403	34	87	30	81
Total		1554	5025	996	3110	573	1380

(2b) percentages of improving the New-2 method

Tools	**PCG with PDF**	**PCG with PFGS**	**New2 method**
NOI	100%	65%	39%
NOF	100%	62%	28%

Clearly there are improvements of both standard BFGS and New2 in a bout 35% - 61% NOI and 38% -64% NOF, respectively

Conclusions:

Clearly, self-scaling techniques are very effective in unconstrained optimization algorithms. The two different approaches used in this paper proved to be very effective, especially for high dimension functions.

Clearly, our numerical results indicate that there are improvements of proposed self-scaling techniques over standard DFP and BFGS algorithms.

REFERENCES

[1] Al-Baali. M. (1985)"Descent property and global convergence of Fletcher-Reeves Method with in exact line search", **J. Inst. Maths. Applns S**,121-124.

[2] Al-Bayati (1996)"A New PCG Method for Unconstrained Non-Linear Optimization", **Journal of Abhath Al-yarmouk, Pure sci. and eng**. Vol. (5), No. (1) Pp. 71-92.

[3] Allwright, J. C. (1972)"Improving the Conditioning of Optimal Control Problems using simple methods", **In Recent Mathematical Developments in Control (ed. Bell, O. J.)**, Academic Press.

[4] Axelsson, O. (1974)"On Preconditioning and Convergence acceleration in sparse matrix problems", **CERN Data Handling Division Report,** 74-1.

[5] Boyd, S. and Vandenberhe, L. (2003) **Convex optimization**, Cambridge University Press.

[6] Buckley, A. G.(1978a) "A combined Conjugate Gradient quasi-Newton minimization algorithm", **Mathematical Programming** 15.

[7] Buckley A. G. (1978b) "Extending the relationship between the Conjugate gradient and BFGS algorithm", **Mathematical Programming** 15.

[8] Dixon, L. C. W. (1972) **Non-linear optimization**, London: English Universities Press.

[9] Edwin, K. and Stamslaw, H. (2001)**An Introduction to Optimization**, John Wiley and. Sons Inc .

[10] Fletcher, R. and Revees, C. (1964) "Function Minimization by Conjugate Gradients", **Computer Journal**, Vol.(7).

[11] Fletcher. R. (1987) "Practical Methods of Optimization", John Wiley and Sons Ltd. Second edition.

[12] Fletcher. R. (1993) "An overview of unconstrained optimization", Numerical Analysis Report NA/ 49.

[13] Gill, Ph. E. Murray and Wright, M. (1981) "Practical Optimization", Academic Press, London.

[14] Hu,Y.F. and Story, C. (1990) "On unconstrained Conjugate gradient optimization Method and their Interrelationship", Mathematics Report Number A129.

[15] Liu, Y. and Story, C. (1991) "Efficient generalized Conjugate Gradient algorithms", Part I, theory J. of Optimization theory and Application, Vol. (169), No. (1).

[16] Nielsen, H. Frandsen; P. E. Jonasson; K. and Tingleff, O. (2004) "Unconstrained Optimization", Informatics and Mathematical Modeling , Technical University of Denmark, 3rd Edition, March .

[17] Nocedal, J. (1992) "Theory of Algorithms for Unconstrained Optimization", Acta Numerica.

[18] Powell, M. J. D. (1977) "Restart procedure for the Conjugate Gradient method", Math. Programming 12.