A new type of optimization method based on conjugate directions

Pan Xin
Science School
Tianjin University of Technology and Education (TUTE)
Tianjin, China
e-mail: panxin1943@sina.com

***Abstract*—A new type of optimization method based on conjugate directions is proposed in this paper. It can be proved that this type of method has quadratic termination property without exact line search. The new method requires only the storage of 4 vectors such that it is suitable for large scale problems. Numerical experiences show that the new method is effective.**

Keywords-optimization; large scale problems; conjugate directions; quadratic termination property

I. INTRODUCTION

Optimization is an important tool in decision science, analysis of physical systems and control engineering. Conjugate gradient (CG) methods are some of the most useful algorithms for unconstrained optimization problem

$$\min f(x), x \in R^n \tag{1}$$

by a sequence of line searches

$$x_{k+1} = x_k + \alpha_k d_k \tag{2}$$

from a user supplied estimate x_1. The step length α_k can be obtained by the line search methods.

Conjugate gradient algorithms are widely used for the two notable properties: First, they need not store any matrix, thus are suitable for large scale problems; Second, when exact line search (E.L.S) is taken, CG algorithms have the quadratic termination property which reflects the efficiency of the algorithms.

The search directions of CG methods are generated as follows

$$d_{k+1} = -g_{k+1} + \beta_k d_k \tag{3}$$

where $g_{k+1} = \nabla f(x_{k+1})$, $d_1 = -g_1$,and β_k is generated by the definition of conjugated directions

$$d_k{}^T G_{k+1} d_{k+1} = 0, k = 1, 2, ... \tag{4}$$

where G_{k+1} is the Hessian of f at the point x_{k+1}.

Various formulas for β_k are given for different CG algorithms[1][2][3],

$$\beta_k = \frac{g_{k+1}^T g_{k+1}}{g_k^T g_k} \tag{5}$$

$$\beta_k = \frac{g_{k+1}^T (g_{k+1} - g_k)}{g_k^T g_k} \tag{6}$$

$$\beta_k = \frac{g_{k+1}^T (g_{k+1} - g_k)}{d_k^T (g_{k+1} - g_k)} \tag{7}$$

In practice an exact line search (E.L.S) is not usually possible for it requires too many evaluations of objective function f and possibly the gradient ∇f. More practical strategies perform an inexact line search (I.L.S) to identify a step length. However, for a general quadratic function, CG methods can't guarantee global convergence when perform I.L.S[4]. Even for a positive quadratic function, when perform I.L.S, CG methods will lose the quadratic termination property.

Example 1: Consider a 2-dimensional positive quadratic function

$$f(x) = \frac{1}{2} x^T A x + b^T x,$$

which gradient is $g(x) = Ax + b$. Suppose that $\{d_1, d_2\}$ is a set of nonzero conjugate vectors.

Case 1: Given a starting point x_1, let us generate the sequence $\{x_k^*\}$ by Eq.(2), where α_k is obtained by E.L.S, given explicitly by

$$\alpha_k = -\frac{g_k^T d_k}{d_k^T A d_k} \tag{8}$$

Thus

$x_2^* = x_1 + \alpha_1^* d_1$, $x_3^* = x_2 + \alpha_2^* d_2$, $g(x_2^*)^T d_1 = 0$, $g(x_3^*)^T d_2 = 0$.

Since $g(x)$ is linear and d_1, d_2 are conjugate, we have

$$g(x_3^*)^T d_1 = \left(g(x_2^*) + \alpha_2^* A d_2\right)^T d_1 = g(x_2^*)^T d_1 = 0$$

So $g(x_3^*) = 0$ and x_3^* is the minimizer of the function $f(x)$. This illustrates the conjugate directions are good search directions when perform E.L.S.

Case 2: Starting from the same x_1, let us perform I.L.S along the search direction d_1 to get the step length $\alpha_1 \neq \alpha_1^*$, and improved $x_2 = x_1 + \alpha_1 d_1$. Then starting from the same x_2 perform E.L.S. along d_2 to get the step length $\overline{\alpha_2}$ and $x_3 = x_2 + \overline{\alpha_2} d_2$. By the property of E.L.S., we have $g(x_3)^T d_2 = 0$. But x_3 is not the minimizer of $f(x)$, for

$$\begin{aligned} g(x_3)^T d_1 &= \left(g(x_2) + \overline{\alpha_2} A d_2\right)^T d_1 = \left(g(x_2^*) + A(x_2 - x_2^*)\right)^T d_1 \\ &= d_1^T A (x_2 - x_2^*) \neq 0 \end{aligned} \tag{9}$$

Case 2 illustrates that when perform I.L.S, CG methods will lose the quadratic termination property. However, some

Published by Atlantis Press, Paris, France.

modification can be taken on the second search direction. Let

$$p_2 = d_2 + \gamma d_1 \tag{11}$$

be the search direction in the second step, and $\hat{\alpha}$ be the step length, $\hat{x_3} = x_2 + \hat{\alpha} p_2$, then

$$\begin{aligned} g(\hat{x_3})^T d_1 &= g(x_2)^T d_1 + \hat{\alpha} d_1{}^T A(d_2 + \gamma d_1) \\ &= d_1{}^T A(x_2 - x_2^*) + \hat{\alpha}\gamma d_1{}^T A d_1 \end{aligned} \tag{12}$$

If we correctly select $\hat{\alpha}$ and γ, we can get $g(\hat{x_3})^T d_2 = 0$ and $g(\hat{x_3})^T d_1 = 0$, and $\hat{x_3}$ is the minimizer. This illustrates that when perform I.L.S, if some modification is taken on the current conjugate direction, we will get a better search direction. The new method in this paper is based on this truth.

II. Derivation of the New Method

Consider *n*-dimensional positive quadratic function

$$f(x) = \frac{1}{2} x^T A x + b^T x + c \tag{13}$$

Suppose p_k is the *k*th search direction, x_{k+1} is the point after the *k*th line search. The Taylor expansion of $f(x)$ can be writen as

$$f(x) = f(x_{k+1}) + g_{k+1}{}^T p + \frac{1}{2} p^T G_{k+1} p \triangleq m(p) \tag{14}$$

where $p = x - x_{k+1}$ $G_{k+1} = A$.

The best search direction in k+1th step is $p_{k+1}^* = \arg\min m(p)$. It is difficult to get the best direction in R^n space if the Hessian is unknown. We can take the idea of CG methods that uses linear combination of two linear independent vectors to approximate p_{k+1}^*. Let accessorial vector d_{k+1} conjugate with p_k,thus $d_{k+1}{}^T A p_k = 0$,then from the hint of Example 1,the search direction in k+1th step can be written as $p_{k+1} = a_1 p_k + a_2 d_{k+1}$. d_{k+1} can be generated form FRCG or PRCG as

$$d_{k+1} = -g_{k+1} + \beta_k p_k . \tag{15}$$

Obviously, if $a_1 = 0$, p_{k+1} is conjugate gradient direction. We named the new method Modified Conjugate Method (MCG). The key to obtain MCG search direction is the selection of combination coefficients a_1 and a_2, which is similar to the idea of Yuan [5].

Let $U = [p_k, d_{k+1}]$, $y = [a_1, a_2]^T$. Then

$$p_{k+1} = Uy \tag{16}$$

$$\begin{aligned} m(p_{k+1}) &= f(x_{k+1}) + g_{k+1}{}^T p_{k+1} + \frac{1}{2} p_{k+1}{}^T G_{k+1} p_{k+1} \\ &= f(x_{k+1}) + (U^T g_{k+1})^T y + \frac{1}{2} y^T U^T G_{k+1} U y \triangleq \varphi(y) \end{aligned} \tag{17}$$

Let $\hat{g} = U^T g_{k+1}$, $\hat{G} = U^T G_{k+1} U$. Then we can rewrite

$$\varphi(y) = f(x_{k+1}) + \hat{g}^T y + \frac{1}{2} y^T \hat{G} y \tag{18}$$

If $y^* = \arg\min_{y \in R^2} \varphi(y)$ is obtained,we can get the minimizer of $m(p)$ on the manifold $\{p = a_1 p_k + a_2 d_{k+1} | a_1 \in R, a_2 \in R\}$ to approximate p_{k+1}^*, written as $p_{k+1} = Uy^*$. p_{k+1} is the search direction of MCG.

If G_{k+1} is positive,then $\hat{G}$ is also positive. So we can get

$$p_{k+1} = Uy^* \tag{19}$$

where $y^* = -\hat{G}^{-1}\hat{g}$. If $\hat{G}$ is not positive,we can take

$$p_{k+1} = -sign(g_{k+1}^T d_{k+1}) d_{k+1} \tag{20}$$

The algorithm is as follows.

Step 1: Given x_k,tolerance ε,set $k = 1$;

Step 2: Calculate $g_k = \nabla f(x_k)$,if $\|g_k\| \le \varepsilon$,stop and x_k is the minimizer; Otherwise, go to step 3;

Step 3: If $k = 1$,set $p_k = -g_k$,go to step 6;Otherwise, set $d_k = -g_k + \beta_k p_{k-1}$,where $\beta_k = \frac{g_k^T (g_k - g_{k-1})}{g_{k-1}^T g_{k-1}}$;

Step 4: Calculate $s = p_{k-1}{}^T G_k p_{k-1}$, $t = d_k{}^T G_k d_k$;

Step 5: If $t \le 0$,then $p_k = -sign(d_k{}^T g_k) d_k$;

Else if $s \le 0$,then $p_k = -sign(p_{k-1}{}^T g_k) p_{k-1}$;

Otherwise $p_k = \frac{p_{k-1}{}^T g_k}{s} p_{k-1} + \frac{d_k{}^T g_k}{t} d_k$,go to step 6;

Step 6: Search α_k which satisfies Wolfe condition：

$$f(x_k + \alpha_k p_k) \le f(x_k) + c_1 \alpha_k \nabla f(x_k)^T p_k$$

$$\nabla f(x_k + \alpha_k p_k)^T p_k \ge c_2 \nabla f(x_k)^T p_k \qquad 0 < c_1 < c_2 < 1$$

Step 7: Set $x_{k+1} = x_k + \alpha_k p_k$,k=k+1,go to step 2.

Remark: Usually, $\alpha = 1$ is always tried first whether it satisfies Wolfe condition when performing E.L.S. That is true in MCG.

III. Properties of MCG

Theorem 1 The search directions generated by MCG are descent directions.

Proof: For the direction generated by Eq.(20),

$$p_{k+1}^T g_{k+1} = -sign(g_{k+1}^T d_{k+1}) g_{k+1}^T d_{k+1} < 0$$

For the direction generated by Eq.(19),

$$\begin{aligned} p_{k+1}{}^T g_{k+1} &= -y^{*T} U^T g_{k+1} \\ &= -g_{k+1}{}^T U \hat{G}^{-1} U^T g_{k+1} \\ &= -\left(\frac{\left(g_{k+1}{}^T p_k\right)^2}{p_k{}^T G_{k+1} p_k} + \frac{\left(g_{k+1}{}^T d_{k+1}\right)^2}{d_{k+1}{}^T G_{k+1} d_{k+1}} \right) < 0 \end{aligned}$$

So the search directions generated by Eq.(19) and Eq.(20) are descent directions. ■

Published by Atlantis Press, Paris, France.

Lemma 1 Suppose that MCG is used to solve n-dimensional positive quadratic problem defined by Eq.(13), the iterative points are $x_1,\cdots,x_i,\cdots$, and search directions generated by MCG are $p_1,p_2,\cdots,p_i,\cdots$, then $\{p_2,p_3,\cdots,p_{n+1}\}$ is a set of nonzero conjugate vectors and for $i\geq 3$, x_i is the minimizer of the objective function on the manifold $\{x_{i-1}+\alpha p_{i-1}|\alpha\in R\}$.

Proof: Since the objective function is positive quadratic, form Eq.(18) and Eq.(19), we can conclude that when using MCG, the step length α_k will be 1, for $k\geq 2$.

For $k\geq 2$, $p_k=U_k y_k$,where

$y_k=\arg\min_{y\in R^2} f(x_k+U_k y)$,

We have $U_k{}^T g(x_k+U_k y_k)=0$,i.e. $U_k{}^T g_{k+1}=0$

Thus $p_k{}^T g_{k+1}=y_k{}^T U_k{}^T g_{k+1}=0$ .It shows that x_{k+1} is the minimizer on $\{x_k+\alpha p_k|\alpha\in R\}$ which means that x_{k+1} is obtained by performing E.L.S along p_k from x_k.

$p_{k+1}=U_{k+1}y_{k+1}$,where

$y_{k+1}=\arg\min_{y\in R^2} f(x_{k+1}+U_{k+1}y)$

$$=\begin{pmatrix}\frac{p_k{}^T g_{k+1}}{s}\\ \frac{d_{k+1}{}^T g_{k+1}}{t}\end{pmatrix}=\begin{pmatrix}0\\ \frac{d_{k+1}{}^T g_{k+1}}{t}\end{pmatrix}$$

It shows that $p_{k+1}\| d_{k+1}$.

We can conclude that $\{p_2,d_3,\cdots,d_{n+1}\}$ is a set of nonzero conjugate vectors from Eq.(15) and $p_{k+1}\| d_{k+1}$ .Therefore $\{p_2,p_3,\cdots,p_{n+1}\}$ is a set of nonzero conjugate vectors.■

Lemma 2 Suppose that $f(x)$ is a n-dimensional positive quadratic function, $\{d_1,d_2,\cdots,d_n\}$ is a set of nonzero conjugate vectors. Let $x_0\in R^n$ be any starting point, if E.L.S. is performed along $d_1,d_2,\cdots,d_n$ to get the sequence $x_1,x_2,\cdots,x_n$, then x_n is the minimizer of $f(x)$ [6].

Theorem 2 MCG has the quadratic termination property.

Proof: From Lemma 1, we know that for positive quadratic function, MCG equals to the conjugate gradient methods starting from the second iterative point which performing E.L.S.. Combined with Lemma 2, it can be concluded that MCG has the quadratic termination property. ■

MCG can be view as a kind of Iterated-subspace minimization methods [7]. The frame work of Iterated-subspace minimization methods (ISM) is as follows.

Step 1: If the *k*th iterative point x_k satisfies the tolerance,stop and output x_k ; Otherwise, continue;

Step 2: Construct $Z_k\in R^{n\times t_k}$ and solve t_k -dimensional sub-problem $y_k=\arg\min_{y\in R^{t_k}} f(x_k+Z_k y)$;

Step 3: $x_{k+1}=x_k+Z_k y_k$, k=k+1, go to step 1.

MCG is a kind of special Iterated-subspace minimization method in which $Z_k=[p_k,d_k]$, $t_k=2$ and Newton method is used to solve the 2-dimensional sub-problems. So the convergence of MCG is based on the convergence of Iterated-subspace minimization methods, which shown as follows [8].

Theorem 3 Suppose that $f(x)$ is a differentiable function, the level set $L=\{x\in R^n\,|\,f(x)\leq f(x_0)\}$ is bounded, $\nabla f(x)$ is Lipschitz-continuous, and for some $\varepsilon>0$, $0<\alpha\leq\beta<1$, the sequence of points generated by the ISM satisfies Goldstein condition

$$f(x_k)+\beta g(x_k)^T Z_k d_z^k \leq f(x_{k+1})\leq f(x_k)+\alpha g(x_k)^T Z_k d_z^k \tag{21}$$

and

$$\frac{-g(x_k)^T Z_k d_z^k}{\left\|Z_k{}^T g(x_k)\right\|_2\left\|d_z^k\right\|_2}\geq\varepsilon \tag{22}$$

where d_z^k is the solution of subproblem $\min_{d_z\in R^{t_k}} f(x_k+Z_k d_z)$, then the ISM algorithm globally converges to a stationary point for problem(1) from any starting point.

Conn [7] proofed that, if negative gradient vector is in the subspace spanned by Z_k and when performing I.L.S. with Wolfe condition in sub-problems, ISM will satisfies Eq.(21) and Eq.(22), thus convergence can be guaranteed. Obviously MCG satisfies the conditions above.

IV. IMPLEMENTATION OF MCG METHOD

The key to implement MCG method is to calculate $s=p_{k-1}{}^T G_k p_{k-1}$ and $t=d_k{}^T G_k d_k$. For the Hessian G_k is not easy to calculate and store, it is reasonable to calculate *s* and *t* approximately. We can use the approximation:

$$G_k p_{k-1}\approx\frac{g(x_k+hp_{k-1})-g(x_k)}{h} \tag{23}$$

$$G_k d_k\approx\frac{g(x_k+hd_k)-g(x_k)}{h} \tag{24}$$

for some small differencing interval h. $h=10^{-8}$ is fairly close to optimal interval value[6]. Then *s* and *t* can be calculated. The algorithm using this approximation is named MCG1. To reduce evaluations of gradient, another approximation can be used:

$$G_k p_{k-1}\approx\frac{g(x_k)-g(x_{k-1})}{\alpha_{k-1}} \tag{25}$$

where α_{k-1} is step length. The algorithm using this approximation is named MCG2.

Published by Atlantis Press, Paris, France.

The algorithms MCG1, MCG2 and PRCG are implemented by Matlab software with the same stopping criterion $\|g(x_k)\|_2 < 10^{-5}$. The test functions are as follows.

TF1：Extended Rosen-brock

$$f(x) = \sum_{i=1}^{n/2} 100(x_{2i} - x_{2i-1}^2)^2 + (1 - x_{2i-1})^2$$

$$x_0 = [-1.2, 1, \cdots, -1.2, 1]^T,\ f(x^*) = 0$$

TF2：Tridia

$$f(x) = \sum_{i=2}^{n} i(2x_i - x_{i-1})^2$$

$$x_0 = (1, 1, \cdots, 1)$$

TF3：Power

$$f(x) = \sum_{i=1}^{n} i x_i^2,$$

$$x_0 = [1, 1, \cdots, 1]^T,\ f(x^*) = 0$$

TF4：Extended Beale

$$f(x) = \sum_{i=1}^{n/2} \begin{Bmatrix} [1.5 - x_{2i-1}(1 - x_{2i})]^2 \\ +[2.25 - x_{2i-1}(1 - x_{2i}^2)]^2 \\ +[2.625 - x_{2i-1}(1 - x_{2i}^3)]^2 \end{Bmatrix},$$

$$x_0 = [1, 1, \cdots, 1]^T$$

TF5：Nondia

$$f(x) = \sum_{i=2}^{n} 100(x_1 - x_i^2)^2 + (1 - x_i)^2$$

$$x_0 = [-1, -1, \cdots, -1]^T,\ f(x^*) = 0$$

The numerical results of MCG1, MCG2 and PRCG are shown in TABLE I. The results are in form of the number of function calls 'Nf', the number of gradient calls 'Ng', the number of iterations 'NI', the central processor unit time 'CPU', the norm of gradient at minimizer 'Norm(g)' and the minimum of the function 'Fvl'. The results show that both MCG1 and MCG2 outperform PRCG. Either MCG1 or MCG2 needs no more than 4 *n*-dimensional vectors to implement. Thus, they are suitable for large scale problems such as 1000-dimensional problems or even higher dimensional problems. Comparing with MCG2, MCG1 needs less function calls but more gradient calls. Therefore, when evaluation of the objective function is much easier than its gradient, MCG2 is the priority option.

V. CONCLUSIONS

The new optimization method named MCG presented in this paper can be view as a kind of Iterated-subspace minimization methods, which global convergence is guaranteed. It has the quadratic termination property without E.L.S. Numerical experiments show that MCG is more efficient and precise than PRCG. Furthermore, MCG is also suitable for large scale problems.

ACKNOWLEDGMENT

This work is supported by NSFC (No. 50975203, No. 51075298). The foundations' support is greatly appreciated.

REFERENCES

[1] Fletcher R. and Reeves C.M. Function Minimization by Conjugate Gradients[J]. Computer Journal, 1964, 7 : 149-154

[2] Sorenson H. W. Comparision of some conjugate direction procedures for function minimization [J]. Journal of the Franklin Institute 1969, 288 :421-441

[3] Liu Y., Storey C. Efficient generalized conjugate gradient algorithms[J]. Journal of optimization theory and applications, 1991,69: 129-153

[4] Dai Y.and Yuan Y. Nonlinear Conjugate Gradient Methods [M]. Shanghai: Shanghai Science and Technology Press, 2001

[5] Yuan Y. and Stoer J. A subspace study on conjugate algorithms[J]. ZAMM Z. angew. Math. Meth. 1995, 75(11):69-77

[6] Nocedal J. and Wright S. J. Numerical Optimization[M]. Springer Series in Operations Research. New York: Springer-Verlag, 1999

[7] [onn A. R., Gould N. L. M., and Sartenaer A.,etal On Interated-Subspace minimization methods for nonlinear optimization[A]. L. Adams and L. Nazareth, Linear and Nonlinear conjugate gradient related methods[C]. AMS-IMS-SIAM. 1996:50-78

[8] Dennis J. E. and Schnabel R.B. Numerical methods for unconstrained optimization and nonlinear equations[M]. Prentice-Hall Englewood Cliffs, NJ, 1983.

TABLE I. Performance of MCG1, MCG2 and PRCG

Function	Dimension	Algorithm	Nf	Ng	NI	CPU	Norm(g)	Fvl
TF1	100	MCG1	76	107	23	0.047	2.62E-07	4.51E-17
	100	MCG2	79	92	26	0.047	2.05E-07	2.59E-17
	100	PRCG	180	154	38	0.047	8.04E-07	5.40E-13
	1000	MCG1	76	107	23	0.078	4.70E-07	1.36E-16
	1000	MCG2	79	92	26	0.078	7.65E-07	3.59E-16
	1000	PRCG	180	154	38	0.109	2.54E-06	5.39E-12
TF2	100	MCG1	155	308	77	0.063	7.09E-06	1.17E-13
	100	MCG2	155	232	77	0.062	7.52E-06	1.22E-13
	100	PRCG	300	292	120	0.562	7.41E-06	5.49E-13
	1000	MCG1	573	1144	286	0.469	9.70E-06	5.31E-14

	1000	MCG2	573	859	286	0.407	9.80E-06	5.31E-14
	1000	PRCG	1192	1171	517	2.406	8.96E-06	3.64E-13
TF3	100	MCG1	107	212	53	0.047	9.98E-06	9.47E-13
	100	MCG2	107	160	53	0.046	9.98E-06	9.47E-13
	100	PRCG	223	212	86	0.062	8.93E-06	1.76E-12
	1000	MCG1	357	712	178	0.297	9.77E-06	2.75E-13
	1000	MCG2	357	535	178	0.25	9.77E-06	2.75E-13
	1000	PRCG	784	765	334	0.421	7.08E-06	6.17E-13
TF4	100	MCG1	25	39	9	0.032	1.08E-06	1.60E-12
	100	MCG2	26	23	9	0.031	2.28E-07	5.34E-16
	100	PRCG	69	59	17	0.047	4.11E-06	9.50E-12
	1000	MCG1	25	39	9	0.078	3.38E-06	1.64E-11
	1000	MCG2	26	33	9	0.078	5.39E-07	3.08E-15
	1000	PRCG	72	61	18	0.125	9.14E-06	9.24E-11
TF5	100	MCG1	51	75	16	0.046	1.65E-07	1.44E-18
	100	MCG2	64	62	16	0.062	9.63E-08	3.69E-19
	100	PRCG	103	79	26	0.079	9.52E-06	2.36E-11
	1000	MCG1	59	83	17	0.094	2.06E-07	1.24E-19
	1000	MCG2	70	90	24	0.093	6.81E-08	2.79E-20
	1000	PRCG	98	79	27	0.125	8.48E-06	1.79E-16

Published by Atlantis Press, Paris, France.