Yokohama Publishers
ISSN 1349-8169 ONLINE JOURNAL

Pacific Journal of Optimization
Vol. 15, No. 3, pp. 457 – 475, 2019

A LIMITED MEMORY CLASS OF CONJUGATE GRADIENT METHODS

Masoud Fatemi

Abstract: We combine the good properties of the linear conjugate gradient algorithm using an optimization problem, and introduce a new limited memory class of nonlinear conjugate gradient methods. This new class contains Dai-Liao family as a subclass. Using this idea, we propose a bound for the optimal Dai-Liao parameter. The global convergence of the new method is investigated under mild assumptions. The numerical comparing results indicate that the new method is efficient and competitive with CG-DESCENT.

Key words: *conjugate gradient method, Dai-Liao family, limited memory, unconstrained optimization, line search*

Mathematics Subject Classification: *90C52, 65K05, 49M37, 26B25*

1 Introduction

We consider the following unconstrained optimization problem,

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

where f is a smooth function. Conjugate gradient algorithms are a class of efficient algorithms for solving (1.1), specially, when n is large [1, 2, 5, 6, 8–11, 15, 16, 18, 19]. This class of methods were originally invented by Hestenes and Stiefel [16] for solving a symmetric and positive definite linear system of equations and then extended by many authors to handle general optimization problems [3, 14]. In a conjugate gradient algorithm, a sequence of iterates, x_{k+1}, is generated by the following scheme,

$$x_{k+1} = x_k + \alpha_k d_k, \tag{1.2}$$

where d_k is a search direction satisfying,

$$d_{k+1} = -g_{k+1} + \beta_k d_k, \quad d_0 = -g_0. \tag{1.3}$$

and, $\alpha_k > 0$ is a step-length to fulfill the Wolfe conditions,

$$f(x_{k+1}) - f(x_k) \le c_1 \alpha_k g_k^T d_k, \tag{1.4}$$

$$g_{k+1}^T d_k \ge c_2 g_k^T d_k, \tag{1.5}$$

where $0 < c_1 < c_2 < 1$ are some arbitrary constants and $g_k := \nabla f(x_k)$.

In the linear conjugate gradient method, the scalars β_k and α_k are so chosen that the following properties hold:

© 2019 Yokohama Publishers

(i) The search direction d_{k+1} is a sufficient descent direction, namely, there exists a scalar $c > 0$ such that

$$g_{k+1}^T d_{k+1} \leq -c \parallel g_{k+1} \parallel^2 . \tag{1.6}$$

(ii) The following conjugacy condition holds:

$$d_{k+1}^T y_{k-i} = 0, \tag{1.7}$$

for $i = 0, \ldots, k$.

(iii) The gradient vector g_{k+1} is orthogonal to the Krylov subspace of degree k, namely,

$$g_{k+1}^T d_{k-i} = 0, \tag{1.8}$$

for $i = 0, \ldots, k$.

It is shown in the linear conjugate gradient theory that the method (1.2) and (1.3) with a search direction satisfying (i)-(iii) terminates in finite iterations. Unfortunately, these properties can not be guaranteed for the non-linear case.

Recently, Hager and Zhang [15] introduced an efficient non-linear conjugate gradient method which is a subclass of the Dai-Liao family corresponding to

$$\beta_k^{DL} = \frac{g_{k+1}^T y_k - \tau g_{k+1}^T s_k}{y_k^T d_k}, \tag{1.9}$$

where $\tau = \lambda_k \frac{\|y_k\|^2}{s_k^T y_k}$. The authors showed the global convergence of a truncated version of their method similar to PRP$^+$ of Gilbert and Nocedal [12] under mild assumptions. Numerical results showed that the method is efficient and robust. Now, an implementation of the method called CG-DESCENT is available from the Hager's homepage.

In order to design an efficient non-linear conjugate gradient method, we combine (i)-(iii) and introduce the following optimization problem

$$\min_{\beta_k} g_{k+1}^T d_{k+1} + M \sum_{i=0}^{m} \left[(g_{k+2}^T s_{k-i})^2 + (d_{k+1}^T y_{k-i})^2 \right], \tag{1.10}$$

where M is the penalty or weight function, and m is an arbitrary constant controlling the memory size. It is easy to see that the first term in (1.10) contains the information about (i) and the second term contains the information about (ii) and (iii). A large value of M increases the chance of d_{k+1} and g_{k+2} to satisfy (1.7) and (1.8), respectively, and a small one increases the effect of sufficient descent property.

Here, we obtain a new expression for β_k by solving (1.10). We show that the optimal Dai-Liao parameter τ must be somewhere in the interval $(0, 1)$. Furthermore, we introduce the two weight functions M_1, corresponding to a generalization of CG-DESCENT, and M_2, and analyze the global convergence of the related algorithms.

The paper is organized as follow: In Section 2, we determine β_k by solving (1.10). Introducing the weight functions M_1 and M_2 with the related expressions for β_k is the subject of Section 3. The global convergence of the new methods is investigated in Section 4 and the numerical result is reported in Section 5. Finally, conclusions and discussions are made in last section.

2 The New Limited Memory Class

In this section, we introduce a new family of conjugate gradient methods by solving (1.10).

In order to solve (1.10), we should replace g_{k+2} by some of its appropriate estimation, because it is not available in the current iteration.

Here, we consider the quadratic approximation of the objective function,

$$\Phi(d) = f_{k+1} + g_{k+1}^T d + \frac{1}{2} d^T B_{k+1} d,$$

and take $\nabla\Phi(\alpha_{k+1}d_{k+1})$ as an estimation of g_{k+2}. It is easy to see that

$$\nabla\Phi(\alpha_{k+1}d_{k+1}) = \alpha_{k+1}B_{k+1}d_{k+1} + g_{k+1}. \tag{2.1}$$

Unfortunately, α_{k+1} in (2.1) is not available, because d_{k+1} is unknown. Thus, we modify (2.1) and set

$$g_{k+2} = tB_{k+1}d_{k+1} + g_{k+1}, \tag{2.2}$$

where $t > 0$ is a suitable approximation of α_{k+1}.

Now, using (1.3) and (2.2), the following expression for β_k is obtained by solving (1.10).

$$\begin{aligned}\beta_k = \frac{1}{X}\Bigg[& - g_{k+1}^T d_k + 2Mt^2 \sum_{i=0}^{m} (s_{k-i}^T B_{k+1} g_{k+1})(s_{k-i}^T B_{k+1} d_k) \\ & - 2Mt \sum_{i=0}^{m} (s_{k-i}^T g_{k+1})(s_{k-i}^T B_{k+1} d_k) \\ & + 2M \sum_{i=0}^{m} (y_{k-i}^T g_{k+1})(y_{k-i}^T d_k)\Bigg],\end{aligned} \tag{2.3}$$

where

$$X = 2Mt^2 \sum_{i=0}^{m} (s_{k-i}^T B_{k+1} d_k)^2 + 2M \sum_{i=0}^{m} (y_{k-i}^T d_k)^2.$$

To simpilify (2.3), we consider the following assumption as Yuan and Stoer [20].

(A1) The approximation matrix B_{k+1} satisfy the extended quasi newton equation

$$B_{k+1} s_{k-i} = y_{k-i},$$

for $i = 0, \ldots, m$.

Using (A1), we have

$$\begin{aligned}\beta_k = & \frac{-1}{2MY(1+t^2)} g_{k+1}^T d_k + \frac{1}{Y} \sum_{i=0}^{m} (y_{k-i}^T g_{k+1})(y_{k-i}^T d_k) \\ & - \frac{t}{Y(1+t^2)} \sum_{i=0}^{m} (s_{k-i}^T g_{k+1})(y_{k-i}^T d_k),\end{aligned} \tag{2.4}$$

where

$$Y = \sum_{i=0}^{m} (y_{k-i}^T d_k)^2.$$

We note that (A1) is hold for strongly convex quadratic functions with exact line search and is called the hereditary property.

Remark 2.1. If $f(x)$ is a strongly convex quadratic function and the line search is exact, then, β_k converts to $\beta_k^{HS} = \frac{y_k^T g_{k+1}}{y_k^T d_k}$ due to (ii) and (iii), and it is exactly the standard conjugate gradient method.

In the reminder of this section, we analyze β_k when M approaches to infinity.

Indeed, the best choice for the weight function M is $M = \infty$, because it increases the probability of satisfying (ii) and (iii).

Approaching M to infinity, we obtain

$$\beta_k = \frac{1}{Y}\sum_{i=0}^{m}(y_{k-i}^T g_{k+1})(y_{k-i}^T d_k) - \frac{t}{Y(1+t^2)}\sum_{i=0}^{m}(s_{k-i}^T g_{k+1})(y_{k-i}^T d_k), \tag{2.5}$$

It is easy to see that (2.5) is a generalization of the Dai-Liao family. More exactly, setting $m = 0$,

$$\beta_k = \frac{g_{k+1}^T y_k}{y_k^T d_k} - \frac{t}{1+t^2}\frac{g_{k+1}^T s_k}{y_k^T d_k},$$

which equals to β_k^{DL} with

$$\tau = \frac{t}{1+t^2}.$$

Note that, τ belongs to $(0, 1)$. As a consequence, we see that the optimal value of the Dai-Liao parameter τ must be somewhere in $(0, 1)$.

Unfortunately, using standard Wolfe line search, it is impossible to ensure the descent property of the search direction d_{k+1} equipped with β_k defined by (2.5). In other words, by approaching M to infinity, we lost the information about the descent property of d_{k+1}. This difficulty can be overcome if we choose an appropriate weight function M. In next section, we follow this idea.

3 The Weight Functions and Descent Directions

In this section, we intended to introduce some suitable weight functions and to show that the corresponding β_k's ensure the sufficient descent property (i).

The first weight function is

$$M_1 = \frac{\gamma_1}{2z\sum_{i=0}^{m} \parallel y_{k-i} \parallel^2}, \tag{3.1}$$

where

$$z = \max(m+1, \frac{\gamma_2 \parallel s_k \parallel \sum_{i=1}^{m} \parallel s_{k-i} \parallel}{\sum_{i=0}^{m} \parallel y_{k-i} \parallel^2}). \tag{3.2}$$

Our motivations behind this choice are firstly to design a limited memory conjugate gradient method which is a generalization of CG-DESCENT, and secondly, to design a method which satisfies the sufficient descent condition independent of the line search procedure. As we will show later, β_k in (2.4) equipped with $M = M_1$ reduces to CG-DESCENT when $m = 0$, and so, the method inherits the good properties of the CG-DESCENT method. Here, we do not claim that $M = M_1$ is the only possible choice, but, at least our numerical results confirm the effectiveness of this choice.

The following lemma indicates that d_{k+1} with β_k as in (2.4) and $M = M_1$ is a sufficient descent direction.

Lemma 3.1. *Assume the method (1.2) and (1.3) with the standard Wolfe line search, where β_k is defined in (2.4) and $M = M_1$. Then, for some positive scalers γ_1 and γ_2 satisfying $\frac{\gamma_1}{4} + \frac{\gamma_1}{2\gamma_2} < 1$,*

$$g_{k+1}^T d_{k+1} \leq -(1 - \frac{\gamma_1}{4} - \frac{\gamma_1}{2\gamma_2}) \parallel g_{k+1} \parallel^2, \tag{3.3}$$

whenever

$$t = \frac{\gamma_1 y_k^T s_k}{z \sum_{i=0}^m \parallel y_{k-i} \parallel^2}. \tag{3.4}$$

Proof. We rewrite β_k with $M = M_1$ in (2.4) as

$$\begin{aligned} \beta_k = \frac{1}{Y} \sum_{i=0}^m (y_{k-i}^T g_{k+1})(y_{k-i}^T d_k) - \frac{A}{Y} g_{k+1}^T d_k \\ - \frac{B}{Y} \sum_{i=1}^m (s_{k-i}^T g_{k+1})(y_{k-i}^T d_k), \end{aligned} \tag{3.5}$$

where

$$A = \frac{1 + 2M_1 t(y_k^T s_k)}{2M_1(1 + t^2)},$$

and

$$B = \frac{t}{1 + t^2}.$$

It is easy to see

$$\begin{aligned} g_{k+1}^T d_{k+1} = - \parallel g_{k+1} \parallel^2 + \frac{1}{\Gamma} \sum_{i=0}^m (y_{k-i}^T g_{k+1})(y_{k-i}^T s_k)(g_{k+1}^T s_k) \\ - \frac{A}{\Gamma} (g_{k+1}^T s_k)^2 - \frac{B}{\Gamma} \sum_{i=1}^m (s_{k-i}^T g_{k+1})(y_{k-i}^T s_k)(g_{k+1}^T s_k), \end{aligned} \tag{3.6}$$

where

$$\Gamma = \sum_{i=0}^m (y_{k-i}^T s_k)^2.$$

Using inequality

$$ab \leq \frac{t'}{4} a^2 + \frac{1}{t'} b^2,$$

where a, b and t' are positive scalers, we have

$$\begin{aligned} g_{k+1}^T d_{k+1} \leq - \parallel g_{k+1} \parallel^2 + \frac{t'}{4\Gamma} \sum_{i=0}^m (y_{k-i}^T g_{k+1})^2 (y_{k-i}^T s_k)^2 \\ + \frac{m+1}{\Gamma t'} (g_{k+1}^T s_k)^2 - \frac{A}{\Gamma} (g_{k+1}^T s_k)^2 \\ + \frac{B}{\Gamma} \sum_{i=1}^m |s_{k-i}^T g_{k+1}| |y_{k-i}^T s_k| |g_{k+1}^T s_k|. \end{aligned}$$

Let $t' = \frac{m+1}{A}$, we have

$$g_{k+1}^T d_{k+1} \leq - \parallel g_{k+1} \parallel^2 + \frac{m+1}{4A} \sum_{i=0}^{m} (y_{k-i}^T g_{k+1})^2 + \frac{B}{2y_k^T s_k} \sum_{i=1}^{m} |s_{k-i}^T g_{k+1}||g_{k+1}^T s_k|. \tag{3.7}$$

Here, we use the inequalities

$$\frac{(y_{k-i}^T s_k)^2}{\Gamma} \leq 1, \tag{3.8}$$

and

$$2|y_k^T s_k||y_{k-i}^T s_k| \leq (y_k^T s_k)^2 + (y_{k-i}^T s_k)^2 \leq \Gamma, \tag{3.9}$$

to obtain (3.7). Finally, using the Cauchy-Schwarz inequality, we have

$$g_{k+1}^T d_{k+1} \leq -\Bigg[1 - \frac{m+1}{4A} \sum_{i=0}^{m} \parallel y_{k-i} \parallel^2 - \frac{B \parallel s_k \parallel \sum_{i=1}^{m} \parallel s_{k-i} \parallel}{2y_k^T s_k}\Bigg] \parallel g_{k+1} \parallel^2 . \tag{3.10}$$

Note that, using (3.1) and (3.2),

$$\frac{m+1}{4A} \leq \frac{M_1(1+t^2)z}{2(1+2M_1 t(y_k^T s_k))} = \frac{\gamma_1}{4\sum_{i=0}^{m} \parallel y_{k-i} \parallel^2}, \tag{3.11}$$

and

$$\begin{aligned} \frac{B \parallel s_k \parallel \sum_{i=1}^{m} \parallel s_{k-i} \parallel}{2y_k^T s_k} &\leq \frac{t \parallel s_k \parallel \sum_{i=1}^{m} \parallel s_{k-i} \parallel}{2y_k^T s_k} \\ &\leq \frac{\gamma_1 \parallel s_k \parallel \sum_{i=1}^{m} \parallel s_{k-i} \parallel}{2z\sum_{i=0}^{m} \parallel y_{k-i} \parallel^2} \\ &\leq \frac{\gamma_1}{2\gamma_2}. \end{aligned} \tag{3.12}$$

Now, the proof is completed using (3.10), (3.11) and (3.12). □

If we substitute (3.1) in (3.5), then

$$\beta_k^1 = \frac{1}{Y} \sum_{i=0}^{m} (y_{k-i}^T g_{k+1})(y_{k-i}^T d_k) - \frac{z}{Y\gamma_1} \sum_{i=0}^{m} \parallel y_{k-i} \parallel^2 (g_{k+1}^T d_k) - \frac{t}{Y(1+t^2)} \sum_{i=1}^{m} (s_{k-i}^T g_{k+1})(y_{k-i}^T d_k), \tag{3.13}$$

where t is defined in (3.4). It is easy to see that setting $m = 0$, β_k^1 converts to the Hager and Zhang choice of β_k^{HZ} with $\lambda_k = \frac{1}{\gamma_1}$, see [15]. Therefore, we can see the method (1.2) and (1.3) with $\beta_k = \beta_k^1$ as a generalization of CG-DESCENT method.

Now, let us to consider the second weight function as

$$M_2 = \frac{2\gamma_3}{(m+1)(1+t^2)\sum_{i=0}^{m} \parallel y_{k-i} \parallel^2}. \tag{3.14}$$

Our motivations behind this choice is firstly to design a new weight parameter with the faster growing rate than M_1, and secondly, to design a method with guaranteeing the sufficient descent property. We reach to the first goal by finding an appropriate interval of t. More exactly, we now try to pass some information about M_1 to t and introduce a new weight parameter M_2 with the faster growing rate than M_1. To see this situation, compare (3.6) and (3.17), closely. In the end of this section, we will show that M_2 grows faster than M_1 to infinity.

The following lemma indicates that for a suitable choice of t, d_{k+1} with β_k as in (2.4) and $M = M_2$ is a sufficient descent direction.

Lemma 3.2. *Assume the method (1.2) and (1.3) with the standard Wolfe line search, where β_k is defined in (2.4) and $M = M_2$. Then, for some positive scalers γ_3 and γ_4 satisfying $\gamma_3 + \gamma_4 < 1$,*

$$g_{k+1}^T d_{k+1} \leq -(1 - \gamma_3 - \gamma_4) \parallel g_{k+1} \parallel^2, \tag{3.15}$$

whenever

$$t \leq \frac{2\gamma_4 (y_k^T s_k)}{\parallel s_k \parallel \sum_{i=0}^{m} \parallel s_{k-i} \parallel}. \tag{3.16}$$

Proof. The proof is essentially similar to the proof of Lemma 3.1.

Using (2.4), we have

$$\begin{aligned} g_{k+1}^T d_{k+1} = &- \parallel g_{k+1} \parallel^2 + \frac{1}{\Gamma} \sum_{i=0}^{m} (y_{k-i}^T g_{k+1})(y_{k-i}^T s_k)(g_{k+1}^T s_k) \\ &- \frac{1}{2M_2\Gamma(1+t^2)} (g_{k+1}^T s_k)^2 \\ &- \frac{t}{\Gamma(1+t^2)} \sum_{i=0}^{m} (s_{k-i}^T g_{k+1})(y_{k-i}^T s_k)(g_{k+1}^T s_k), \end{aligned} \tag{3.17}$$

where

$$\Gamma = \sum_{i=0}^{m} (y_{k-i}^T s_k)^2.$$

Similar to the proof of Lemma 3.1, we reach to

$$\begin{aligned} g_{k+1}^T d_{k+1} \leq &- \parallel g_{k+1} \parallel^2 + \frac{t'}{4\Gamma} \sum_{i=0}^{m} (y_{k-i}^T g_{k+1})^2 (y_{k-i}^T s_k)^2 \\ &+ \frac{m+1}{\Gamma t'} (g_{k+1}^T s_k)^2 - \frac{1}{2M_2\Gamma(1+t^2)} (g_{k+1}^T s_k)^2 \\ &+ \frac{t}{\Gamma(1+t^2)} \sum_{i=0}^{m} |s_{k-i}^T g_{k+1}||y_{k-i}^T s_k||g_{k+1}^T s_k|. \end{aligned}$$

Let $t' = 2M_2(m+1)(1+t^2)$, we have

$$\begin{aligned} g_{k+1}^T d_{k+1} \leq &- \parallel g_{k+1} \parallel^2 + \frac{M_2(m+1)(1+t^2)}{2} \sum_{i=0}^{m} (y_{k-i}^T g_{k+1})^2 \\ &+ \frac{t}{2(1+t^2) y_k^T s_k} \sum_{i=0}^{m} |s_{k-i}^T g_{k+1}||g_{k+1}^T s_k|. \end{aligned} \tag{3.18}$$

Finally, the Cauchy-Schwarz inequality implies

$$g_{k+1}^T d_{k+1} \leq -\Bigg[1-\frac{M_2(m+1)(1+t^2)}{2}\sum_{i=0}^{m} \parallel y_{k-i} \parallel^2 - \frac{t}{2(1+t^2)s_k^T y_k} \parallel s_k \parallel \sum_{i=0}^{m} \parallel s_{k-i} \parallel \Bigg] \parallel g_{k+1} \parallel^2 . \tag{3.19}$$

Now, the proof is completed using (3.14), (3.16) and (3.19). □

Note that, substituting M_2 in (2.4), we have

$$\beta_k^2 = \frac{1}{Y}\sum_{i=0}^{m}(y_{k-i}^T g_{k+1})(y_{k-i}^T d_k) - \frac{m+1}{4Y\gamma_3}\sum_{i=0}^{m} \parallel y_{k-i} \parallel^2 (g_{k+1}^T d_k) - \frac{t}{Y(1+t^2)}\sum_{i=0}^{m}(s_{k-i}^T g_{k+1})(y_{k-i}^T d_k), \tag{3.20}$$

where t is a suitable approximation of α_{k+1} satisfying (3.16).

Lemma 3.2 indicates that there is a degree of freedom in choosing t. In other words, we are free to choose a value for t between zero and

$$\frac{2\gamma_4(y_k^T s_k)}{\parallel s_k \parallel \sum_{i=0}^{m} \parallel s_{k-i} \parallel}.$$

Since t is an approximation of α_{k+1}, it is reasonable to take

$$t = \min\left(p_k, \frac{2\gamma_4(y_k^T s_k)}{\parallel s_k \parallel \sum_{i=0}^{m} \parallel s_{k-i} \parallel}\right), \tag{3.21}$$

where p_k is some approximation of α_{k+1}. In our numerical tests, we take $p_k = \alpha_k$.

A closer look at M_1 and M_2 reveals that, for $m \geq 1$, M_2 can grow faster than M_1 to infinity if we consider the following assumption:

(A2) The gradient g is Lipschitz continuous; namely, there exists a constant $L > 0$ such that

$$\parallel \nabla f(x) - \nabla f(y) \parallel \leq L \parallel x - y \parallel, \quad x, y \in \mathbb{R}^n.$$

It is easy to see that

$$M_1 = O\left(\frac{1}{z\sum_{i=0}^{m} \parallel y_{k-i} \parallel^2}\right), \tag{3.22}$$

and

$$M_2 = O\left(\frac{1}{\sum_{i=0}^{m} \parallel y_{k-i} \parallel^2}\right). \tag{3.23}$$

Note that, we use (3.16), (A2) and the fact that

$$t \leq \frac{2\gamma_4(y_k^T s_k)}{\parallel s_k \parallel \sum_{i=0}^{m} \parallel s_{k-i} \parallel} \leq 2\gamma_4 L \tag{3.24}$$

to obtain (3.23).

For $m \geq 1$, it is possible that z to become a large value. Thus, we can expect a larger value for M_2 than M_1.

As we explained, a large value for M is desirable. As a consequence, we can expect the better result of the method (1.2) and (1.3) with $\beta_k = \beta_k^2$ than the method with $\beta_k = \beta_k^1$. The numerical results of Section 5 confirm our claim.

4 Global Convergence

We now investigate the global convergence of the method (1.2) and (1.3) for both $\beta_k = \beta_k^1$ and $\beta_k = \beta_k^2$. We also assume that the step length α_k satisfy the standard Wolfe conditions (1.4) and (1.5).

The following standard assumptions are considered in this section.

(A3) $f(x)$ is differentiable and bounded below.

(A4) The generated sequence of iterates, x_k, is bounded.

The global convergence of descent methods with standard Wolfe line search relies essentially on the following Zoutendijk condition.

Lemma 4.1. *suppose that A2-A4 holds. consider any descent method of the form (1.2) where α_k is determined by standard Wolfe line search. Then we have that*

$$\sum_{k=1}^{\infty} \frac{(g_k^T d_k)^2}{\| d_k \|^2} < \infty. \tag{4.1}$$

Our global convergence analysis is similar to that of Hager and Zhang in [15]. Here, we consider the following modification version of β_k^i:

$$\beta_k^{(i)} = \max(\beta_k^i, \chi_k), \tag{4.2}$$

where $i \in \{1, 2\}$ and χ_k is a real valued function with the following properties:

p1. $|\chi_k| \| d_k \|$ is bounded.

p2. For some $0 < \epsilon < 1$,

$$\chi_k \leq \frac{\epsilon \| g_{k+1} \|^2}{g_{k+1}^T d_k},$$

whenever, $g_{k+1}^T d_k > 0$.

Note that, p1 ensures that the search direction d_k is bounded and p2 ensures that the sufficient descent property (1.6) holds.

Lemma 4.2. *Suppose the method (1.2) and (1.3) with the choice of $\beta_k = \beta_k^{(i)}$ where i is a fixed value belonging to $\{1, 2\}$. Moreover, assume that the standard Wolfe line search conditions (1.4) and (1.5) are used, then*

$$g_{k+1}^T d_{k+1} \leq -\max(1 - \epsilon, \xi_i) \| g_{k+1} \|^2, \tag{4.3}$$

where

$$\xi_i = \begin{cases} 1 - \frac{\gamma_1}{2} - \frac{\gamma_1}{2\gamma_2}, & i=1; \\ 1 - \gamma_3 - \gamma_4, & i=2. \end{cases}$$

Proof. If $\beta_k = \beta_k^i$, then (3.3) and (3.15) imply that

$$g_{k+1}^T d_{k+1} \leq -\xi_i \parallel g_{k+1} \parallel^2 .$$

If $\beta_k = \chi_k$ and $g_{k+1}^T d_k < 0$, then our previous analysis and the fact that $\beta_k^i \leq \chi_k$ imply that

$$\begin{aligned} g_{k+1}^T d_{k+1} &= - \parallel g_{k+1} \parallel^2 + \chi_k g_{k+1}^T d_k \\ &\leq - \parallel g_{k+1} \parallel^2 + \beta_k^i g_{k+1}^T d_k \\ &\leq -\xi_i \parallel g_{k+1} \parallel^2 . \end{aligned}$$

If $\beta_k = \chi_k$ and $g_{k+1}^T d_k > 0$, then property p2 implies

$$\begin{aligned} g_{k+1}^T d_{k+1} &= - \parallel g_{k+1} \parallel^2 + \chi_k g_{k+1}^T d_k \\ &\leq -(1-\epsilon) \parallel g_{k+1} \parallel^2 . \end{aligned}$$

Now, the proof is completed. □

The following lemma is analogue of Lemma 4.3 in [4]

Lemma 4.3. *Suppose A1-A4 holds, then for method (1.2) and (1.3) with $\beta_k = \beta_k^{(i)}$ where i is a fixed value belonging to $\{1, 2\}$, and a line search satisfying standard Wolfe conditions, we have*

$$\sum_{k=1}^{\infty} \parallel u_k - u_{k-1} \parallel^2 < \infty,$$

where $u_k = \frac{d_k}{\|d_k\|}$, whenever $\inf \parallel g_k \parallel \neq 0$.

Proof. The proof is basically similar to Lemma 4.3 in [4]. We let

$$z_k^{(1)} = \max(\beta_k^i - \chi_k, 0),$$

and

$$z_k^{(2)} = \chi_k.$$

It is easy to see that $\beta_k = z_k^{(1)} + z_k^{(2)}$. Let

$$w_k = \frac{-g_k + z_{k-1}^{(2)} d_{k-1}}{\parallel d_k \parallel},$$

and

$$\delta_k = \frac{z_{k-1}^{(1)} \parallel d_{k-1} \parallel}{\parallel d_k \parallel}.$$

Following the statements of the proof of Lemma 4.3 in [4], we reach to

$$\parallel u_k - u_{k-1} \parallel \leq 2 \parallel w_k \parallel,$$

Since we assumed that $|\chi_k| \parallel d_k \parallel$ is bounded, there exists a constant $\epsilon > 0$ such that

$$\parallel -g_k + z_{k-1}^{(2)} d_{k-1} \parallel \leq \parallel g_k \parallel + |\chi_{k-1}| \parallel d_{k-1} \parallel \leq \epsilon.$$

Thus,

$$\parallel u_k - u_{k-1} \parallel \leq \frac{2\epsilon}{\parallel d_k \parallel}.$$

Now, the proof is completed using (4.3) and the Zoutendijk condition. □

There are some special choices of χ_k in literatures. For example, the Hager and Zhang choice of

$$\chi_k = \frac{-1}{\| d_k \| \min(\eta, \| g_k \|)}, \tag{4.4}$$

and the Dai and Kon choice of

$$\chi_k = \eta \frac{g_{k+1}^T d_k}{\| d_k \|^2}.$$

We now state the main result.

Theorem 4.4. *Suppose that assumptions A1-A4 hold. If the method (1.2) and (1.3) with $\beta_k = \beta_k^i$ where i is a fixed value belonging to $\{1, 2\}$, is implemented on f and the standard Wolfe line search conditions (1.4) and (1.5) are used, then*

$$\liminf_{k\to\infty} \| g_k \| = 0.$$

Proof. Assume that there exists η_1 such that $\| g_k \| > \eta_1$ for all k.

If there exists a subsequence k_j such that $\beta_{k_j} = \chi_{k_j}$, then, using p1, we have for some $\epsilon > 0$,

$$\begin{aligned} \| d_{k_j+1} \| &= \| -g_{k_j+1} + \beta_{k_j} d_{k_j} \| \\ &\leq \| g_{k_j+1} \| + |\chi_{k_j}| \| d_{k_j} \| \\ &\leq \epsilon. \end{aligned}$$

This bound for d_{k_j+1} and (4.3) yield a contradiction using Zoutendijk condition.

We now assume that $\beta_k = \beta_k^i$ for sufficiently large k. Following the statements of the proof of Theorem 3.2 in [15], we only address the changes in the parts of the proof.

Part I. (A bound for β_k) Using inequalities

$$y_k^T d_k \geq (1 - c_2) \max(1 - \epsilon, \xi_i) \eta_1^2, \tag{4.5}$$

and

$$\frac{|g_{k+1}^T d_k|}{|y_k^T d_k|} \leq \max(\frac{c_2}{1 - c_2}, 1), \tag{4.6}$$

see [15], we show that there exists a constant $C > 0$ such that

$$|\beta_k| \leq C \sum_{i=0}^{m} \| s_{k-i} \| . \tag{4.7}$$

It is easy to see using (3.13) and (3.20) that

$$\begin{aligned} |\beta_k| \leq \frac{1}{Y} \sum_{i=0}^{m} |y_{k-i}^T g_{k+1}| |y_{k-i}^T d_k| + \gamma_5 \frac{m + 1 + \gamma_6^k}{Y} \sum_{i=0}^{m} \| y_{k-i} \|^2 |g_{k+1}^T d_k| \\ + \frac{t}{Y(1 + t^2)} \sum_{i=0}^{m} |s_{k-i}^T g_{k+1}| |y_{k-i}^T d_k|, \end{aligned}$$

where

$$\gamma_5 = \max(\frac{1}{4\gamma_3}, \frac{1}{\gamma_1}),$$

and

$$\gamma_6^k = \frac{\gamma_2 \parallel s_k \parallel \sum_{i=1}^{m} \parallel s_{k-i} \parallel}{\sum_{i=0}^{m} \parallel y_{k-i} \parallel^2}.$$

Now, (3.8), (3.9) and the fact that $\frac{t}{1+t^2} < 1$ imply

$$|\beta_k| \leq \frac{1}{y_k^T d_k}\Big[\frac{1}{2}\sum_{i=0}^{m}|y_{k-i}^T g_{k+1}| + \gamma_5(m+1+\gamma_6^k)\sum_{i=0}^{m} \parallel y_{k-i} \parallel^2 \frac{|g_{k+1}^T d_k|}{y_k^T d_k} + \frac{1}{2}\sum_{i=0}^{m}|s_{k-i}^T g_{k+1}|\Big].$$

Using cauchy-Schwarz inequality,

$$|\beta_k| \leq \frac{1}{y_k^T d_k}\Big[\frac{\eta_2}{2}\sum_{i=0}^{m} \parallel y_{k-i} \parallel + \gamma_5(m+1)\sum_{i=0}^{m} \parallel y_{k-i} \parallel^2 \frac{|g_{k+1}^T d_k|}{y_k^T d_k} + \gamma_5\gamma_2 \parallel s_k \parallel \sum_{i=1}^{m} \parallel s_{k-i} \parallel \frac{|g_{k+1}^T d_k|}{y_k^T d_k} + \frac{\eta_2}{2}\sum_{i=0}^{m} \parallel s_{k-i} \parallel \Big], \tag{4.8}$$

where η_2 is an upper bound on $\parallel g_k \parallel$.

Now, it is easy to see from (4.8) that using (A2), (A4), (4.5) and (4.6), there exist a constant $C > 0$ such that (4.7) holds.

Part II. (A bound on steps) Following the statements of the proof of Part II of Theorem 3.2 in [15], we let Δ to be a positive integer such that

$$\Delta \geq 4(m+1)CD + m, \tag{4.9}$$

where D is the diameter of $\{x_k \mid k \in \mathbb{N}\}$. Choose k_0 large enough that

$$\sum_{i \geq k_0} \parallel u_{i+1} - u_i \parallel^2 \leq \frac{1}{4\Delta}.$$

Note that, k_0 is well defined due to Lemma 4.3. Following the statements of the proof of part II of Theorem 3.2 in [15], we have

$$\sum_{j=k}^{l-1} \parallel s_j \parallel \leq 2D, \tag{4.10}$$

when $l > k > k_0$ and $l - k \leq \Delta$.

Part III. (A bound on the directions) Using (4.7), we have

$$\parallel d_l \parallel^2 \leq (\parallel g_l \parallel + |\beta_{l-1}| \parallel d_{l-1} \parallel)^2 \leq 2\eta_2^2 + 2C^2(\sum_{i=0}^{m} \parallel s_{l-i-1} \parallel)^2 \parallel d_{l-1} \parallel^2 .$$

Let

$$S_j = 2C^2(\sum_{i=0}^{m} \parallel s_{j-i} \parallel)^2.$$

Thus, for $l > k_0 + m$,

$$\parallel d_l \parallel^2 \leq 2\eta_2^2\Big(\sum_{i=k_0+m+1}^{l}\prod_{j=i}^{l-1} S_j\Big) + \Big(\prod_{j=k_0+m}^{l-1} S_j\Big) \parallel d_{k_0+m} \parallel . \tag{4.11}$$

Let us focus on the product of $\Delta - m$ consecutive S_j, where $k \geq k_0 + m$:

$$\begin{aligned}
\prod_{j=k}^{k+\Delta-m-1} S_j &= \prod_{j=k}^{k+\Delta-m-1} 2C^2 (\sum_{i=0}^{m} \| s_{j-i} \|)^2 \\
&= \Big(\prod_{j=k}^{k+\Delta-m-1} \sqrt{2} C \sum_{i=0}^{m} \| s_{j-i} \| \Big)^2 \\
&\leq \Big(\frac{\sum_{j=k}^{k+\Delta-m-1} (\sqrt{2} C \sum_{i=0}^{m} \| s_{j-i} \|)}{\Delta - m} \Big)^{2(\Delta-m)} \\
&\leq \Big(\frac{m\sqrt{2} C \sum_{j=k-m}^{k+\Delta-m-1} \| s_j \|}{\Delta - m} \Big)^{2(\Delta-m)} \\
&\leq \Big(\frac{2\sqrt{2}(m+1) CD}{\Delta - m} \Big)^{2(\Delta-m)} \\
&\leq (\frac{1}{2})^{\Delta-m} \qquad (4.12)
\end{aligned}$$

The first inequality is the arithmetic geometric mean inequality and the second and third inequalities comes from (4.10) and (4.9). Since (4.12) is independent of l, we can deduce similar to Part III of Theorem 3.2 in [15] that $\| d_l \|$ is bounded independent of $l > k_0$. This is a contradiction using Zoutendijk condition. □

5 Numerical Results

We now investigate the numerical behavior of the two algorithms presented in the previous sections. The first algorithm is based on the method (1.2) and (1.3) with $\beta_k = \beta_k^1$ (based on the weight function M_1) and is called M1Cgm, where m is a specific memory size. The second algorithm is based on the method (1.2) and (1.3) with $\beta_k = \beta_k^2$ (based on the weight function M_2) and t as in (3.21) with $p_k = \alpha_k$. This algorithm is called M2Cgm. In our numerical tests, we consider the fourth versions of the algorithms corresponding to the choices of $m = 0, 1, 3$ and 5. Moreover, we note again that M1Cg0 is actually CG-DESCENT with $\lambda_k = \frac{1}{\gamma_1}$. The reported results of M1Cg0 was obtained by downloading and running CG-DESCENT code obtained from its web page. We compare all versions of the algorithms on unconstrained problems of CUTEr collection [13]. All runs were performed in MATLAB 2007 on a 2.4 Intel Core 2Duo processor computer with 2GB of RAM. The performance profile of Dolan and Moré [7] is used to compare the efficiency of the algorithms. Furthermore, We used the CG-DESCENT line search procedure with the initialization parameters reported in [15] in our implementations. As in the CG-DESCENT, the algorithms terminate if either

$$\| \nabla f(x_k) \|_\infty \leq \max(10^{-6}, 10^{-12} \| \nabla f(x_1) \|_\infty),$$

our the number of iteration exceed 50000. In our algorithms, we used the following initial parameters:

$$\gamma_1 = 1,\ \gamma_2 = 2,\ \gamma_3 = 0.98 \text{ and } \gamma_4 = 0.01.$$

Our rational for these choices was the following: The chosen values $\gamma_1 = 1$ and $\gamma_2 = 2$ ensure that the method converts to CG-DESCENT with the best reported parameter $\lambda_k = 1$, when $m = 0$. It is easy to see that the sufficient descent condition (1.6) holds with $c = \frac{1}{2}$. The chosen values $\gamma_3 = 0.98$ and $\gamma_4 = 0.01$ ensure that the weight function M_2 in (3.14) is as

large as possible. Note that, using (3.24), a small choice of γ_4 implies that the term $1 + t^2$ appeared in the denominator of (3.14) is as small as possible. It is easy to see that the sufficient descent condition (1.6) holds with $c = 0.01$.

In figures 1-4, we compare M1Cgm with M2Cgm for the same values of m. these figures give the performance profile for the number of iteration, the number of function and gradient evaluations. Part (a) of figures 1, 2 and 4 indicate that M2Cg0 have a better performance than M1Cg0 (CG-DESCENT), specially, in the terms of function and gradient evaluations. Their efficiency are approximately the same for number of iterations. In practice, we observed that choosing a good approximation p_k of α_{k+1} can strongly improve the efficiency of M2Cgm. For $m \geq 1$, figures 1-4 also indicate that M2Cgm strongly dominate M1Cgm for all terms. The domination increases by growing m. In the end of Section 3, we claim that for $m \geq 1$, the algorithm based on the weight function $M2$ can produce the better result than the algorithm based on the weight function $M1$. We showed that $M2$ can grow faster than $M1$ to infinity. Figures 1-4 confirm our claim.

In figures 1(d), 3(b) and 4(d), we also show the performance profile of the method when β_k is defined by (2.5) in comparison with the best of all the algorithms, more exactly, M2Cg5 and M1Cg5 . We denote the corresponding method by MCg∞. We also consider the following simple truncation strategy to guarantee the sufficient descent property.

$$\beta_k = \begin{cases} \beta_k^{\infty} \text{as defined by (2.5)}, & g_{k+1}^T d_{k+1} \leq -0.1||g_{k+1}||^2; \\ 0, & \text{O.W.} \end{cases}$$

As the figures indicated, M2Cg5 and M1Cg5 wins MCg∞. It seems the reason of the poor efficiency is due to our truncation strategy. In this case, a suitable truncation strategy needs more investigation.

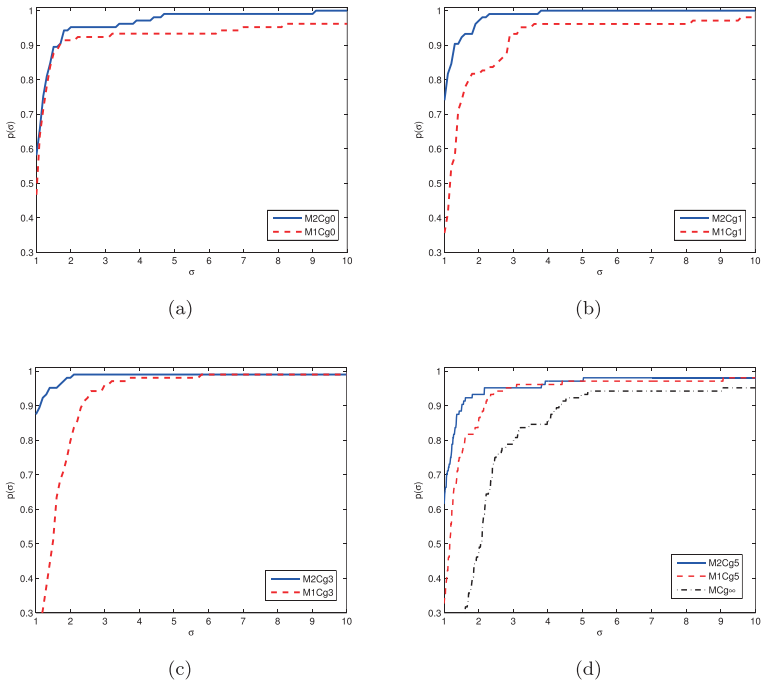
Figures 5 and 6 give the performance profile of M1Cgm and M2Cgm, for $m = 0, 1, 3$ and 5. It is easy to see that the efficiency of the algorithms increase by growing m. This confirms the effectiveness of the memory structure.

6 Conclusions

We have presented a new limited memory class of nonlinear conjugate gradient methods. We showed that this class contains Dai-Liao family as a subclass. As a consequence, we obtained a bound for the optimal Dai-Liao parameter. The global convergence of the new method was investigated under mild assumptions. The numerical comparing results indicated that the new method is efficient and competitive with CG-DESCENT.

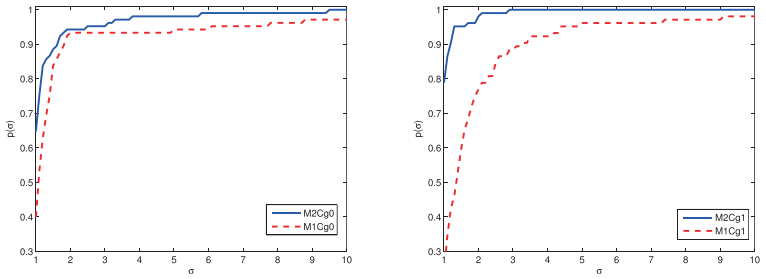
Acknowledgment

The author thank the Research Council of K. N. Toosi University of Technology for supporting this work.



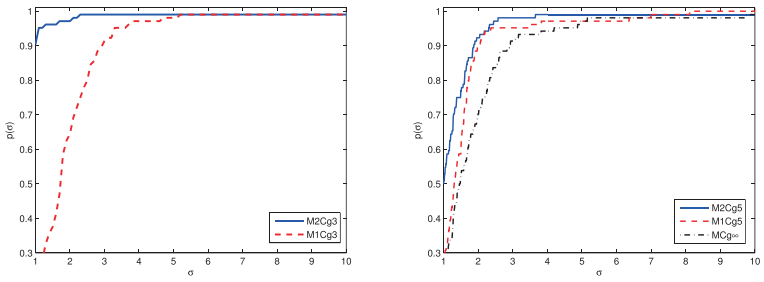
(a) (b) (c) (d)

Figure 1: Iteration performance profile.



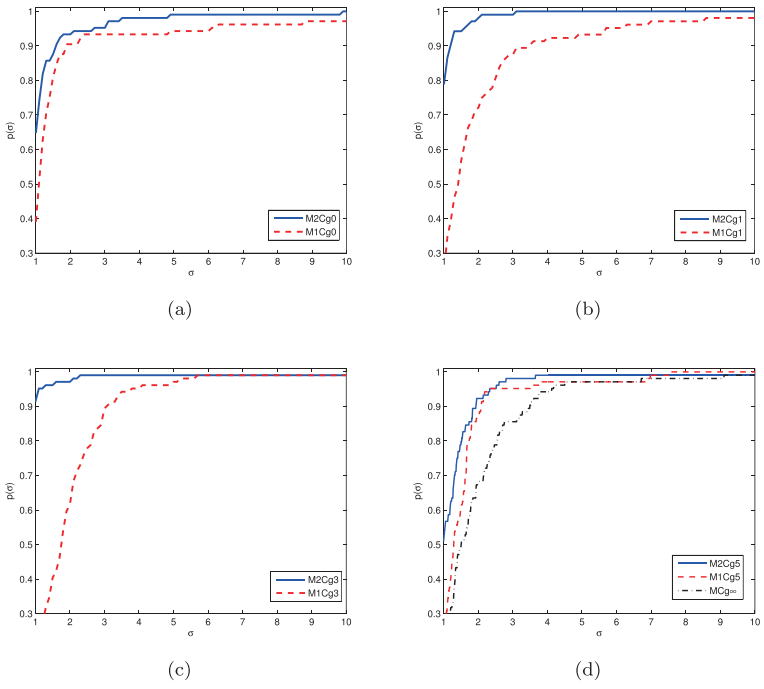
(a) (b)

Figure 2: Number of function evaluation performance profile.



(a) (b)

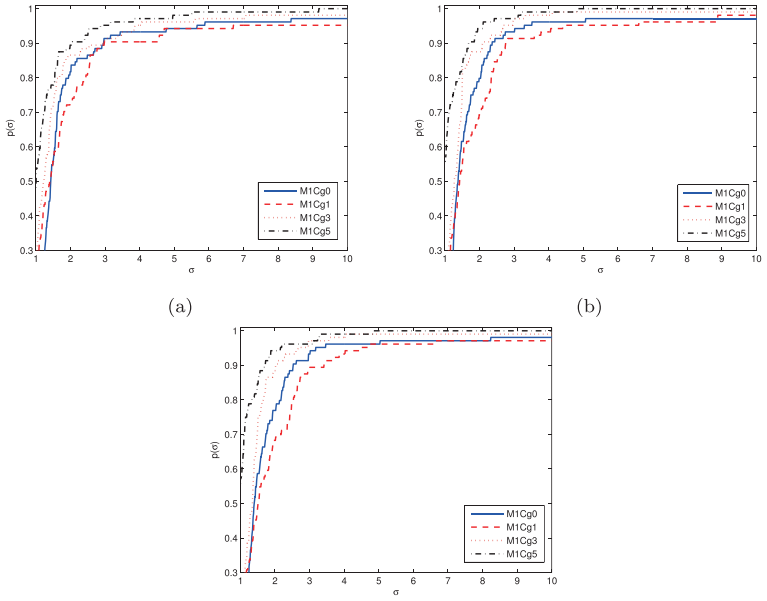
Figure 3: Number of function evaluation performance profile (Continued).



(a) (b)

(c) (d)

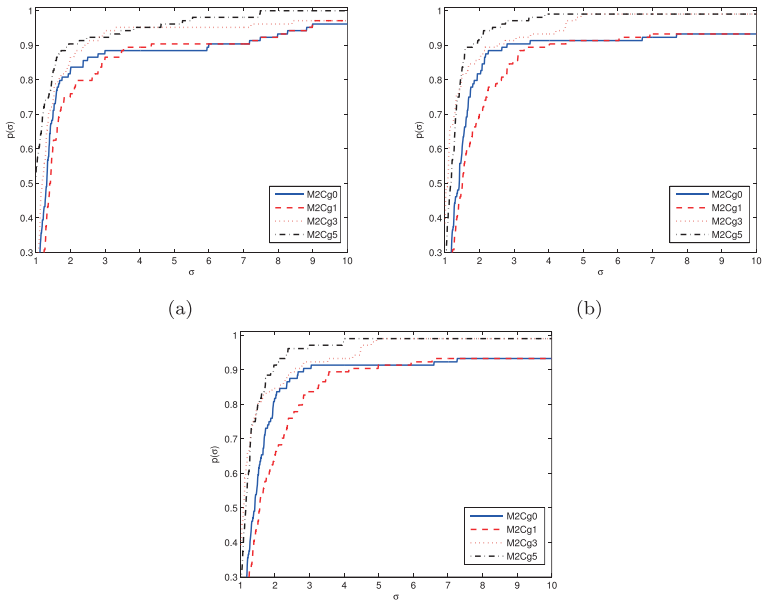
Figure 4: Number of gradient evaluation performance profile.



(a) (b)

(c)

Figure 5: M1Cgm: (a) Iteration performance profile. (b) Number of function evaluation performance profile. (c) Number of gradient evaluation performance profile.



(a) (b)

(c)

Figure 6: M2Cgm: (a) Iteration performance profile. (b) Number of function evaluation performance profile. (c) Number of gradient evaluation performance profile.

References

[1] S. Babaie–Kafaki and M. Fatemi, A modified two-point stepsize gradient algorithm for unconstrained minimization. *Optim. Methods Softw.* 5 (2013) 1040–1050.

[2] S. Babaie-Kafaki, M. Fatemi and N. Mahdavi-Amiri, Two effective hybrid conjugate gradient algorithms based on modified BFGS updates. *Numer. Algorithms.* 3 (2011) 315–331.

[3] Y.H. Dai, *Nonlinear Conjugate Gradient Methods*, Wiley Encyclopedia of Operations Research and Management Science, 2011.

[4] Y.H. Dai and C.X. Kou, A nonlinear conjugate gradient algorithm with an optimal property and an improved wolfe line search. *SIAM J. Optim.* 23 (2013) 296–320.

[5] Y.H. Dai and L.Z. Liao, New conjugacy conditions and related nonlinear conjugate gradient methods. *Appl. Math. Optim.* 43 (2001) 87–101.

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

[7] E.D. Dolan and J.J. Moré, Benchmarking optimization software with performance profile. *Math. Program.* 91 (2002) 201–213

[8] M. Fatemi, An optimal parameter for Dai-Liao family of conjugate gradient methods. *J. Optim. Theory Appl.* 169 (2016) 587–605.

[9] M. Fatemi, A new efficient conjugate gradient method for unconstrained optimization. *J. Comput. Appl. Math.* 300 (2016) 207–216.

[10] M. Fatemi and S. Babaie-Kafaki, Two extensions of the Dai-Liao method with sufficient desent property based on a penalization scheme. *Bull. Comput. Appl. Math.* 4 (2016) 7–19.

[11] R. Fletcher and C.M. Reeves, Function minimization by conjugate gradients. *J. Comput.* 7 (1964) 149–154.

[12] J.C. Gilbert and J. Nocedal, Global convergence properties of conjugate gradient methods for optimization. *SIAM J. Optim.* 2 (1992) 21–42.

[13] N.I. M. Gould, D. Orban and P.L. Toint, CUTEr (and SifDec), A constrained and unconstrained testing environment, revisited. *ACM Trans. Math. Software.* 29 (2003) 373–394.

[14] W.W. Hager and H. Zhang, A survey of nonlinear conjugate gradient methods. *Pac. J. Optim.* 2 (2006) 335–358.

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

[16] M.R. Hestenes and E. Stiefel, *Method of conjugate gradient for solving linear system.* Vol. 49. NBS, 1952.

[17] J.J. Moré and D. J. Thuente, Line search algorithms with gauranteed sufficient decrease. *ACM Trans. Math. Software.* 20 (1994) 286–307.

[18] E. Polak and G. Ribière, Not sur la convergence de méthods directions conjugées. *Revue Franccaise d'Informatique et de Recherche opiérationnelle.* 16 (1969) 35–43.

[19] B.T. Polyak, The conjugate gradient method in extremem problems. *Comput. Math. Math. Phys.* 9 (1969) 4–112.

[20] Y. Yuan and J. Stoer, A subspace study on conjugate gradient algorithms. *ZAMM Z. Angew. Math. Mech.* 75 (1995) 69–77.

Manuscript received 14 November 2015
revised 1 June 2016
accepted for publication 13 July 2016

Masoud Fatemi
Department of Mathematics
K. N. Toosi University of Technology
Tehran, Iran
E-mail address: smfatemi@kntu.ac.ir